2019

Object-Based Supervised Machine Learning Regional-Scale Land-Cover Classification Using High Resolution Remotely Sensed Data

Christopher A. Ramezan
West Virginia University, christopher.ramezan@mail.wvu.edu

Object-Based Supervised Machine Learning Regional-Scale Land-Cover Classification Using High Resolution Remotely Sensed Data

Christopher A. Ramezan

Dissertation submitted to the Eberly College of Arts and Sciences

at West Virginia University

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Geography

Timothy A. Warner, Ph.D., Chair

Jamison Conley, Ph.D.

Gregory Elmes, Ph.D.

Rick Landenberger, Ph.D.

Ramesh Sivanpillai, Ph.D.

Department of Geology and Geography

Morgantown, West Virginia

2019

Keywords: Remote Sensing, GEOBIA, NAIP, LIDAR, High Resolution, Machine Learning, Supervised Classification, Regional-Scale, Land-Cover Mapping

Abstract

Object-Based Supervised Machine Learning Regional-Scale Land-Cover Classification Using High Resolution Remotely Sensed Data

Christopher A. Ramezan

High spatial resolution (HR) (1m – 5m) remotely sensed data in conjunction with supervised machine learning classification are commonly used to construct land-cover classifications. Despite the increasing availability of HR data, most studies investigating HR remotely sensed data and associated classification methods employ relatively small study areas. This work therefore drew on a 2,609 km^2, regional-scale study in northeastern West Virginia, USA, to investigates a number of core aspects of HR land-cover supervised classification using machine learning. Issues explored include training sample selection, cross-validation parameter tuning, the choice of machine learning algorithm, training sample set size, and feature selection. A geographic object-based image analysis (GEOBIA) approach was used. The data comprised National Agricultural Imagery Program (NAIP) orthoimagery and LIDAR-derived rasters. Stratified-statistical-based training sampling methods were found to generate higher classification accuracies than deliberative-based sampling. Subset-based sampling, in which training data is collected from a small geographic subset area within the study site, did not notably decrease the classification accuracy. For the five machine learning algorithms investigated, support vector machines (SVM), random forests (RF), *k*-nearest neighbors (*k*-NN), single-layer perceptron neural networks (NEU), and learning vector quantization (LVQ), increasing the size of the training set typically improved the overall accuracy of the classification. However, RF was consistently more accurate than the other four machine learning algorithms, even when trained from a relatively small training sample set. Recursive feature elimination (RFE), which can be used to reduce the dimensionality of a training set, was found to increase the overall accuracy of both SVM and NEU classification, however the improvement in overall accuracy diminished as sample size increased. RFE resulted in only a small improvement the overall accuracy of RF classification, indicating that RF is generally insensitive to the Hughes Phenomenon. Nevertheless, as feature selection is an optional step in the classification process, and can be discarded if it has a negative effect on classification accuracy, it should be investigated as part of best practice for supervised machine land-cover classification using remotely sensed data.

Dedication

To my parents, Amir and Kathryn Ramezan

Table of Contents

Chapter 1 - Introduction

1. Background

High-spatial resolution (HR) (1m – 5m) earth observation datasets can be used to develop land-cover classifications of the earth's surface to monitor our ever-changing landscape. Over the past few decades, starting with the launch in late 1999 of IKONOS, the first commercially available spatial HR earth observation satellite, HR remotely sensed data has become increasingly available for public use. However, HR remotely sensed data are rarely analyzed on large, regional or national scales. Regional-scale in this context simply refers to a large, multi-county study area, rather than definitions used in disciplines such as landscape ecology or ecology.

Typically, regional- or national-scale land-use/land-cover (LULC) maps are constructed from medium- or coarse-spatial resolution remotely sensed data. For example, the 2011 National Land Cover Dataset (NLCD) is based on medium resolution Landsat data, providing a 30m land cover dataset over the contiguous United States (Homer et al., 2015). Global-scale datasets, such as the Vegetation Continuous Fields (VCF) dataset, constructed from MODIS imagery, typically have an even coarser spatial resolution, such as 250 m or 1 km. While medium or coarse-spatial resolution datasets can be useful, the scale is inappropriate for studies that require finer spatial detail (Li and Shao, 2014), such as urban feature extraction (Benediktsson et al., 2003; Kong et al., 2006; Taubenböck et al., 2010), tree crown mapping (Karlson et al., 2014), forest structural parameter estimation (Galidaki, et al., 2016; Wolter et al., 2009) and small-area site-specific crop management (SSCM) or precision agriculture mapping (Mulla, 2013).

Despite advances in information technology platforms, and increasing availability of HR remotely-sensed datasets, large-scale HR land-cover classifications remain rare (Ma et al., 2017; O'Neil-Dunne et al., 2014). The lack of large-scale (large-scale in this context meaning large geographic areas) land-cover analyses is likely due in part to the relatively large data volume of HR datasets compared to coarser resolution datasets covering the same area. To partially assist with remediating data volume and processing issues with large-scale HR datasets, Geographic Object-Based Image Analysis (GEOBIA) approaches have been suggested for conducting large-scale HR land-cover classifications (Demarchi et al., 2017; Li and Shao, 2014; Lubker and Schaab, 2010).

GEOBIA is a relatively recent paradigm in remote sensing which has several potential advantages over traditional pixel-based approaches when analyzing HR remotely sensed data (Arvor et al., 2013; Blaschke, 2010; Blaschke et al., 2014; Hay and Castilla, 2008; Hay and Blaschke, 2010). GEOBIA approaches can mitigate some of the technical challenges in classifying HR data through the grouping of heterogeneous pixels into discrete image-objects. Rather than having to analyze and classify each pixel individually, segmented groups of pixels, called image-objects, form the base unit of analysis, thus reducing the overall number of data units and in turn reducing the processing demands (Zhang et al., 2007). In addition, GEOBIA approaches can help to reduce the effects of high intra-class spectral variability or "salt and pepper" texture, which is a commonly encountered obstacle in HR classifications of remotely sensed data (Blaschke, 2010).

Object-based analyses of HR remotely sensed data have been conducted for a variety of applications, such as land-use/land-cover mapping (Antonarakis, et al., 2008; Elhadi et al., 2014; Im et al., 2013; Lu and Weng, 2007; Liu and Xia, 2009; Tehrany et al., 2014; Walker and Blaschke, 2006; Zhou et al., 2008) tree-canopy mapping (Chen and Hay, 2011; O'Neil-Dunne et al., 2014;

Machala and Zejdova, 2014), mine-reclaimed land mapping (Maxwell and Warner, 2015), hydrological mapping (Demarchi et al., 2017), bathymetric mapping (Diesing et al., 2014; Diesing et al., 2016; Lacharité et al., 2015), and acoustic remote sensing (Hill et al., 2014), among others. While GEOBIA has become a popular method for analyzing remotely sensed datasets, a majority of basic and applied object-based analyses using HR data reported in the literature are also conducted on geographically small areas (Ma et al., 2017).

Supervised machine learning algorithms are a popular method for constructing land-cover classifications in GEOBIA and remote sensing analyses in general. Supervised machine learning classifiers are mathematical algorithms that use pre-labeled training examples to infer a function, which can be applied to classify new unseen examples. Machine learning has become increasingly popular in a variety of fields, such as biomedical science (Cao et al., 2018) and automotive engineering (Huval et al., 2015). Machine learning methods are particularly attractive to remote sensing scientists due to the typical high data volume and complexity of remotely sensed data.

While many core methodological themes and issues in GEOBIA-based supervised machine learning land-cover classifications such as sample selection, cross-validation, machine learning algorithms, and classification optimization, have been widely examined within the literature, these concepts are rarely, if ever investigated on large, HR regional-area datasets.

This work therefore seeks to investigate methods for developing large area, supervised machine learning regional-scale object-based land-cover classifications of HR remotely sensed data. As traditional remote sensing methods applied to large datasets may be expensive in terms of human and computer effort, this work examines several core HR remote sensing methodological questions, such as sample selection, model cross-validation, supervised machine

learning algorithms, and feature optimization at large, regional scales. In addition, this work provides a practical approach for developing applied regional-scale HR land-cover classifications datasets using an object-based approach.

2. Research Themes and Aims

Through investigation of the development of regional-scale object-based high-resolution land-cover classifications, this dissertation addresses the following research questions:

1. As sample selection is a critical component of conducting supervised machine learning classifications, how do various aspects of sample acquisition processes such as sample size, sample location, and sampling technique affect the accuracy of supervised regional-scale land-cover classifications of HR remotely sensed data?

2. Cross-validation tuning is often used to optimize parameter selection in supervised machine learning classifiers. While a variety of cross-validation tuning methods are commonly used in remote sensing analyses, they are rarely compared. Do different cross-validation tuning methods provide inherent advantages for improving supervised classifier performance of regional-scale HR remotely sensed data?

3. As various supervised machine learning methods are commonly employed for constructing land-cover classifications of remotely sensed data, how do the following supervised machine learning classifiers: Support Vector Machine (SVM), Random Forests (RF), *k*-Nearest Neighbors (k-NN), Neural Networks (NEU), and Learning Vector Quantization (LVQ) perform when constructing regional-scale HR land-cover maps?

4. High dimensional remotely sensed datasets can contain highly correlated or irrelevant bands (or features) that can negatively affect the performance of supervised machine learning classifiers, a problem known as the curse of dimensionality or the Hughes Phenomenon (Hughes, 1968). Automated feature selection approaches have therefore been suggested for optimizing the feature space of training sets, and reducing the data dimensionality. This work investigates if recursive feature elimination improves classifier performance when applied to large, regional-scale HR remotely sensed datasets and using three supervised machine learning classifiers: Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NEU).

3. Objectives and Structure

This dissertation consists of three main chapters, each written as a stand-alone article, with its own experimental design. Additionally, an introduction chapter that explains the overall context of the work, and a conclusion chapter that summarizes and reflects on the findings of the three articles, are included.

Chapter 2 focuses on research question one, examining how the performance of support vector machines (SVM) supervised regional-scale HR land-cover classification is affected by training sets that vary in size, acquisition location, and collection method. In addition, chapter 2 also examines the second question in a comparison of cross-validation tuning methods in a regional-scale land-cover classification of HR remotely sensed data.

Chapter 3 focuses on the first and third research questions by investigating how training samples of varying size acquired from differing geographic regions affect the performance of several different supervised machine learning algorithms when applied to a regional-scale HR land-cover classification.

Chapter 4 focuses on the first and fourth research questions by examining how feature selection techniques such as can be used to improve performance of regional-scale HR supervised classifications trained from varying size training sets.

4. References

Arvor, D., L. Durieux, S. Andrés, and Laporte, M.A. (2013). Advances in Geographic Object-Based Image Analysis with Ontologies: A Review of Main Contributions and Limitations from a Remote Sensing Perspective. *ISPRS Journal of Photogrammetry and Remote Sensing,* 82: 125–137. doi:10.1016/j.isprsjprs.2013.05.003

Antonarakis, A. S., Richards, K. S., Basington, J. (2008). Object-based land cover classification using airborne LiDAR. *Remote Sensing of Environment.* 112(6). 2988-2998. https://doi.org/10.1016/j.rse.2008.02.004.

Benediktsson, J. A., Pesaresi, M., Amason, K. (2003). Classification and feature extraction for remote sensing images from urban areas based on morphological transformations. *IEEE Transactions of Geoscience and Remote Sensing*. 41(9). 1940-1949. DOI: 10.1109/TGRS.2003.814625.

Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65 (1), 2-16. http://dx.doi.org/10.1016/j.isprsjprs.2009.06.004.

Blaschke, T., Hay, G. J., Kelly, M., Lang, S., Hofmann, P., Addink, E., . . . Tiede, D. (2014). Geographic Object-Based Image Analysis – Towards a new paradigm. *ISPRS Journal of Photogrammetry and Remote Sensing,* 87, 180-191. doi:10.1016/j.isprsjprs.2013.09.014.

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z. (2018). Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics & Bioinformatics*. 16(1). 17-32. https://doi.org/10.1016/j.gpb.2017.07.003.

Chen, G., and Hay, G. J. (2011). An airborne lidar sampling strategy to model forest canopy height from Quickbird imagery and GEOBIA. *Remote Sensing of Environment.* 115(6). 1532-1542. https://doi.org/10.1016/j.rse.2011.02.012.

Demarchi, Luca., Bizzi, Simone., Piegay, Herve. (2017). Regional hydromorphological characterization with continuous and automated remote sensing analysis based on VHR imagery and low-resolution LiDAR data. Earth Surface Processes and Landforms. 42(3). 531-551. DOI: 10.1002/esp.4092.

Diesing, M., Green, S. L., Stephens, D., Lark, R. M., Stewart, H. A., and Dove, D. (2014). Mapping seabed sediments: Comparison of manual, geostatistical, object-based image analysis and machine learning approaches. *Continental Shelf Research*. 84 (1). 107-119. https://doi.org/10.1016/j.csr.2014.05.004.

Diesing, M., Mitchell, P., and Stephens, D. (2016). Image-based seabed classification: What can we learn from terrestrial remote sensing? *ICES Journal of Marine Science*. 73(10). 2425-2441.

Elhadi, A., Mutanga, O., Odindi, J., and Abdel-Rahman, E. M. (2014). Land-use/cover classification in a heterogeneous coastal landscape using RapidEye imagery: evaluating the performance of random forest and support vector machine classifiers. *International Journal of Remote Sensing* 35(10). 3440-3458. https://doi.org/10.1080/01431161.2014.903435.

Galidaki, G., Zianis, D., Gitas, I., Radoglou, K., Karathanassi, V., Tsakiri-Strai, M. (2016). Vegetation biomass estimation with remote sensing: focus on forest and other wooded land over the Mediterranean ecosystem. *International Journal of Remote Sensing*. 38(7). 1940-1966. https://doi.org/10.1080/01431161.2016.1266113.

Hay, G. J., and Castilla, G. (2008). Geographic Object-Based Image Analysis (GEOBIA): a new name for a new discipline. In T. Blaschke, S. Lang, G. Hay (EDs.) Object-Based Image Analysis. Springer, Heilderberg, Berlin, NY (2008). 78-85.

Hill, N. A., Lucieer, V., Barrett, N. S., Anderson, T. J., Williams, S. B. (2014). Filling the gaps: Predicting the distribution of temperate reef biota using high resolution biological and acoustic data. *Estuarine, Coastal and Shelf Science.* 147 (20). 137–147. https://doi.org/10.1016/j.ecss.2014.05.019.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K. (2015). Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. Photogrammetric Engineering and Remote Sensing, 81(5), 345-354.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. 14(1). 55-63. DOI: 10.1109/TIT.1968.1054102.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., Cheng-Yue, R., Mujica, F., Coates A., and Ng, A. Y. (2015). An Empirical Evaluation of Deep Learning on Highway Driving. *arXiv: Robotics.* 1-7. https://arxiv.org/pdf/1504.01716v1.pdf.

Im, J., Jensen, J. R., and Hodgson, M. E. (2013). Object-Based Land Cover Classification Using High-Posting-Density LiDAR Data. *GIScience & Remote Sensing*. 45 (2). 209-228. https://doi.org/10.2747/1548.1603.45.2.209.

Karlson, M., Reese, H., and Ostwald, M. (2014). Tree Crown mapping in Managed Woodlands (Parklands) of Semi-Arid West Africa Using WorldView-2 Imagery and Geographic Object Based Image Analysis. *Sensors*. 14(12). 22643-22669. DOI: 10.3390/s141222643.

Kong, Chunfang., Xu, Kai., Wu, Chonglong. (2006). Classification and Extraction of Urban Land-Use Information from High-Resolution Image Based on Object-Multi-features. *Journal of China University of Geosciences*. 17(2). 151-157. https://doi.org/10.1016/S1002-0705(06)60021-6.

Lacharité, M., Metaxas, A., and Lawton P. (2015). Using object-based image analysis to determine seafloor fine-scale features and complexity. *Limnology and Oceanography: Methods*. 13. 553–567. http://doi.wiley.com/10.1002/lom3.10047

Li, Xiaoxiao and Shao, Guofan (2014). Object-Based Land-Cover Mapping with High Resolution Aerial Photography at a County Scale in Midwestern USA. Remote Sensing. 6. 11372-11390. doi:10.3390/rs61111372.

Lu, D., Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28. 823-870.

Lubker, T. and Schaab, G. (2010). A Work-flow design for large-area multilevel GEOBIA: Integrating statistical measure and expert knowledge. In: ISPRS Proceedings (digital) of Geographic Object-Based Image Analysis (GEOBIA 2010), Vol. XXXVIII-4/C7, 29 June – 2 July 2010, Ghent (Belgium); ed. by Addink, E.A. & F.M.B. Van Coillie.

Liu, D., and Xia, F. (2010). Assessing object-based classification: advantages and limitations. *Remote Sensing Letters*. 2010(4). 187-194. https://doi.org/10.1080/01431161003743173.

Ma, L., Li, M., Ma, X., Cheng, K., Du, P., Liu Y. (2017). A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 130(2017). 277-293. https://doi.org/10.1016/j.isprsjprs.2017.06.001.

Machala, M., and Zejova, L. (2014). Forest Mapping through Object-based Image Analysis of Multispectral and LiDAR Aerial Data. *European Journal of Remote Sensing*. 47(1). 117-131. https://doi.org/10.5721/EuJRS20144708.

Maxwell. A. E., and Warner, T. A. (2015). Differentiating mine-reclaimed grasslands from spectrally similar land cover using terrain variables and object-based machine learning classification. *International Journal of Remote Sensing* 36(17). 4384-4410. https://doi.org/10.1080/01431161.2015.1083632.

Mulla, D. J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaning knowledge gaps. *Biosystems Engineering*. 114(4). 358-371. https://doi.org/10.1016/j.biosystemseng.2012.08.009.

O'Neil-Dunne, J., MacFaden, S., Royar, A. (2014). A Versatile, Production-Oriented Approach to High-Resolution Tree-Canopy mapping in Urban and Suburban Landscapes Using GEOBIA and Data Fusion. *Remote Sensing* 6 (12), 12837-12865. https://doi.org/10.3390/rs61212837.

O'Neil-Dunne, J., MacFaden, S., Royar, A., Reis, M., Dubayah, R., Swatantran, A. (2014). An Object-Based Approach to Statewide Land Cover Mapping. *ASPRS 2014 Annual Conference*. Louisville, Kentucky, March 23-28, 2014. 1-6.

Taubenböck, H., Esch, T., Wurm, M., Roth, A., and Dech, S. (2010). Object-based feature extraction using high spatial resolution satellite data of urban areas. *Journal of Spatial Sciences*. 55(1). 117-132. https://doi.org/10.1080/14498596.2010.487854.

Tehrany, M. S., Pradhan, B., and Jebuv, M. N. (2014). A comparative assessment between object and pixel-based classification approaches for land use/land cover mapping using SPOT 5 imagery. *Geocarto International*. 29(4). 351-369. https://doi.org/10.1080/10106049.2013.768300.

Walker, J. S., and Blaschke, T. (2008). Object-based land-cover classification for the Phoenix metropolitan area: optimization vs. transportability. *International Journal of Remote Sensing* 29(7). 2021-2040. https://doi.org/10.1080/01431160701408337.

Wolter, P. T., Townsend, P. A., Sturtevant, B. R. (2009). Estimation of forest structural parameters using 5 and 10 meter SPOT-5 satellite data. *Remote Sensing of Environment*. 113(9). 2019-2036. https://doi.org/10.1016/j.rse.2009.05.009.

Zhang, L., Zhong, Y., Huang, B., Li, P. (2007). A resource limited artificial immune algorithm for supervised classification of multi/hyper-spectral remote sensing image. *International Journal of Remote Sensing*. 28(7), 1665-1686.

Zhou, W., Troy, A., Grove, M. (2008). Object-based Land Cover Classification and Change Analysis in the Baltimore Metropolitan Area using Multitemporal High Resolution Remote Sensing Data. *Sensors*. 8(3). 1613-1636.

Chapter 2

Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification[1]

Abstract

High spatial resolution (1–5 m) remotely sensed datasets are increasingly being used to map land covers over large geographic areas using supervised machine learning algorithms. Although many studies have compared machine learning classification methods, sample selection methods for acquiring training and validation data for machine learning, and cross-validation techniques for tuning classifier parameters are rarely investigated, particularly on large, high spatial resolution datasets. This work, therefore, examines four sample selection methods—simple random, proportional stratified random, disproportional stratified random, and deliberative sampling—as well as three cross-validation tuning approaches—k-fold, leave-one-out, and Monte Carlo methods. In addition, the effect on the accuracy of localizing sample selections to a small geographic subset of the entire area, an approach that is sometimes used to reduce costs associated with training data collection, is investigated. These methods are investigated in the context of support vector machines (SVM) classification and geographic object-based image analysis (GEOBIA), using high spatial resolution National Agricultural Imagery Program (NAIP) orthoimagery and LIDAR-derived rasters, covering a 2,609 km2 regional-scale area in northeastern West Virginia, USA. Stratified-statistical-based sampling methods were found to generate the highest classification accuracy. Using a small number of training samples collected from only a subset of the study area provided a similar level of overall accuracy to a sample of equivalent size collected in a dispersed manner across the entire regional-scale dataset. There were minimal differences in accuracy for the different cross-validation tuning methods. The processing time for Monte Carlo and leave-one-out cross-validation were high, especially with large training sets. For this reason, k-fold cross-validation appears to be a good choice. Classifications trained with samples collected deliberately (i.e., not randomly) were less accurate than classifiers trained from statistical-based samples. This may be due to the high positive spatial autocorrelation in the deliberative training set. Thus, if possible, samples for training should be selected randomly; deliberative samples should be avoided.

[1] Published by MDPI in *Remote Sensing* on 18 January 2019. Available online: https://www.mdpi.com/2072-4292/11/2/185. Ramezan, C. A., Warner, T. A., Maxwell A. E. 2019. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sensing. 11(2): 185. https://doi.org/10.3390/rs11020185.

1. Introduction

With the increasing availability of high spatial resolution (HR) remotely sensed datasets (1–5 m pixels), the routine production of regional-scale HR land-cover maps has become a possibility. However, due to the large area associated with regional-scale HR remote sensing projects, the sample selection for training and assessment can be burdensome. Sampling strategies that are commonly used in remote sensing analyses involving smaller datasets may be unsuitable or impractical for regional-scale HR analyses. This is particularly true if the sampling protocol requires field observations. While much previous remote sensing research has been conducted on supervised classification sample selection methods for training [1–3] and accuracy assessments [4–7], most of these studies examine sampling methods using study sites of limited geographic extent. The limited area of these study sites is typical of classification experiments in general; Ma et al. [8] meta-reviewed over 170 supervised object-based remote sensing analyses and found that an overwhelming majority of geographic object-based image analyses (GEOBIA) studies were conducted on areas smaller than 300 ha.

This work, therefore, investigates a variety of sample selection method techniques for regional-scale land-cover classifications with large, HR remotely sensed datasets. Additionally, as the number of samples is limited in many regional studies, cross-validation for regional HR classification is also explored. Cross-validation is an approach for exploiting training and accuracy assessment samples multiple times and thus potentially improving the reliability of the results. Finally, as the acquisition of widely dispersed samples across a large region may be expensive, a sampling strategy which confines the sample selection to a small geographic subset area is also investigated. This study is conducted in the context of GEOBIA, an approach that has become increasingly popular for analyzing high-resolution remotely sensed data [8,9].

1.1. Background on Sample Selection in Remote Sensing

Samples in remote sensing analyses are typically collected for two purposes: training data for developing classification models and assessment or test data for evaluating the accuracy of the map product. Supervised classifiers, such as machine learning algorithms, use pre-labeled samples to train the classifier, which is then used to assign class labels to the remaining population. As the collection of training data inherently requires sampling, the strategies used for the sample selection must be carefully considered in the context of the characteristics of the dataset, classifier, and study objectives [10]. Although sample selection strategy is widely discussed in the remote sensing literature, there are a variety of opinions on almost every aspect of the sampling process. Nevertheless, there is a consensus that the size [5,7,11,12] and quality [12] of the training sample dataset, as well as the sample selection method used [5], can affect classification and accuracy assessments.

A variety of statistical (e.g., simple random and stratified random) and non-statistical (e.g., deliberative) sample selection methods have been used to collect training and testing samples for remote sensing analyses. Mu et al. [13] separated statistical-based sampling into two categories: spatial and aspatial approaches. Spatial sampling considers the spatial autocorrelation inherent in geographic data, while aspatial methods, which ignore potential spatial autocorrelation, include approaches such as simple random and stratified sampling. Although problems associated with aspatial sampling methods in remote sensing have been noted [14–16], spatial sampling methods can be complex and typically require a priori information about the population, which may be difficult or impractical to collect. While spatial sampling methods have been used in remote sensing analyses, currently they are far less common than aspatial methods and consequently not pursued in this study.

Simple random sampling involves the purely random selection of samples and thus gives a direct estimate of the population parameters. Although random samples for image classification on average will sample each class (or stratum) in proportion to the area of that class in the map, any single random

sample will generally not do so. This can exacerbate the difficulty of dealing with rare classes. Some classifiers, including support vector machines (SVM), are sensitive to imbalanced training data sets, in which some classes are represented by a much smaller number of samples than other classes [17].

Stratified random sampling addresses this problem by forcing the number of samples in each stratum to be proportional to the area of the class. A variant on this method is equalized stratified random sampling, where the number of samples in each stratum is the same, irrespective of their area on the map. Equalized stratified random sampling may not be possible if some classes are so rare that the population of that class is smaller than the desired sample size. In such circumstances, a disproportional stratified random sample may be collected, an approach in which the sizes of the strata are specified by the user and are set to intermediate values between the proportions of the areas of the classes and a simple equalized approach.

One disadvantage of all stratified approaches is that a pre-classification is needed to identify the strata. If the samples are only to be used for an accuracy assessment (and not training), then it is possible to use the classification itself to generate the strata. However, Stehman [18] points out that if multiple classifications are to be compared and the strata are developed from just one of those classifications, the resulting accuracy statistics for the remaining maps need to be modified to account for the differences between the map used to develop the stratification and the map under consideration. Furthermore, unlike random and stratified random sampling, equalized and disproportional stratified random samplings produce samples that are not a direct estimate of the population. Thus, for samples generated with these methods, the accuracy estimates need adjustment to account for the class prior probabilities [19].

Deliberative sampling, in which samples are selected based on a non-random method, are also common in remote sensing. Deliberative sampling is necessary if access limitations or other issues

constrain the sampling. One could hypothesize that deliberative sampling might, under certain circumstances, be more effective for classifier training than random sampling. Deliberative sampling allows for the incorporation of expert knowledge into the sample selection process. For example, samples can be selected to ensure that the variability of each class is well represented. Furthermore, in SVM, only training samples that define the hyperplane separating the classes are used by the classifier. Thus, for SVM, deliberative samples selected to represent potentially spectrally confused areas may be more useful than samples representing the typical class values [20].

If in situ observations are required for sample characterization and the cost of traveling between sites is high, the spatial distribution of samples becomes a central focus. This is particularly a concern for regional-scale HR datasets. While certain innovative methods such as active learning have been proposed to reduce sampling costs [21,22], these methods are complex to implement and are beyond the scope of this study. An alternative is localizing sample selection to a single subset area of the region of interest. Localizing sample selection to a small geographic subset area can be advantageous in large regional-scale analyses for reducing sampling costs, especially if field observations are required.

1.2. Background on Cross-Validation Tuning

A central tenet of accuracy assessment is that the samples used for training should not also be used for evaluation. A similar concern applies to the methods for selecting the user-specified parameters required by most machine learning methods, for example, the number of trees for random forests, sigma and C values for radial basis function kernel support vector machines, and the k-distance for the k-nearest neighbors. The value of these parameters can affect the accuracy of the classification, and thus, optimizing the chosen values (sometimes called tuning) is usually required [23–26]. Tuning is generally empirical, with various values for the parameters systematically evaluated, and the combination of values that generate the highest overall accuracy or kappa coefficient is assumed to be optimal [17,25]. Excluding training samples from the samples used for the evaluation of the candidate

parameter values reduces the likelihood of overtraining and thus improves the generalization of the classifier.

If the overall number of samples is small, a fixed partition of training samples into separate training and tuning samples will further exacerbate the limitations of the small sample size, since each sample is used once and for one purpose only (e.g., training). Cross-validation is an alternative approach to a fixed partition. In cross-validation, multiple partitions are generated, potentially allowing each sample to be used multiple times for multiple purposes, with the overall aim of improving the statistical reliability of the results. Examples of cross-validation methods include k-fold, leave-one-out, and Monte Carlo. Classification parameter tuning via cross-validation has been demonstrated to improve classification accuracy in remote sensing analyses [27]. However, as with any sampling technique, it is important that the overall sample set be representative of the entire data set, otherwise the generalizability of the supervised classifier is unknown [25].

The k-fold cross-validation method involves randomly splitting the sample set into a series of equally sized folds (groups), where k indicates the number of partitions, or folds, the dataset is split into. For example, if a k-value of ten is used, the dataset is split into ten partitions. In this case, nine of the partitions are used for training data, while the remaining one partition is used for test data. The training is repeated ten times, each time using a different partition as the test set and the remaining nine partitions as the training data. The average of the results is then reported [28].

Leave-one-out cross-validation is similar to the k-fold cross-validation except the number of folds is set as the number of samples in the sample set. This approach can be slow with very large sample sets [29].

Monte Carlo validation works on similar principles to k-fold cross-validation except that the folds are randomly chosen with replacement, also called bootstrapping. Thus, the Monte Carlo method may

result in some samples being used for both training and testing data multiple times, or some data not being used at all. Usually, Monte Carlo methods employ a large number of simulations, for example 1,000 or more, and therefore may also be slow [30].

While studies such as by Maxwell et al. [17] and Cracknell and Reading [25] demonstrated the merits of the cross-validation methods such as k-fold cross-validation for parameter tuning, very little attention has been given to examining the different cross-validation methods and their effect on parameter optimization and, by extension, machine learning classification performances.

1.3. Research Questions and Aims

This work examines sample selection and cross-validation parameter tuning on regional-scale land cover classifications using HR remotely sensed data. These issues are explored through the following interlinked research questions:

1. Which training sample selection method results in the highest classification accuracy for a supervised support vector machine (SVM) classification of a regional-scale HR remotely sensed dataset? The methods tested include both statistical (simple random, proportional stratified random, and disproportional stratified random) and non-statistical (deliberative) methods.

2. Which cross-validation method provides the highest classification accuracy? Methods tested are k-fold, leave-one-out, and Monte Carlo.

3. What is the effect on classification accuracy for the different sampling and cross-validation methods when the samples are collected from a small localized region rather than from across the entire study area?

2. Materials and Methods

2.1. Study Area and Data

The study area (Figure 1) lies within the northeastern section of West Virginia, near the borders with Maryland and Pennsylvania. The study area includes the entirety of the Preston County, as well as proportions of the neighboring Monongalia, Taylor, Barbour, and Tucker counties. This region is dominated by Appalachian mixed mesophytic forests [31] and the terrain is mountainous (548–914 m).

Figure 1 - The regional-scale study area

Two remotely sensed datasets were used in this analysis: optical multispectral imagery and light detection and ranging (LIDAR) point cloud data. The optical dataset comprises leaf-on National Agriculture Imagery Program (NAIP) orthoimagery collected primarily during 17–30 July 2011. A very small portion of the NAIP imagery was collected on 10 October 2011. The NAIP imagery consists of four spectral bands (red (590–675 nm), green (500–650 nm), blue (400–580 nm), and near-infrared (NIR) (675–850 nm)), with an 8-bit radiometric resolution and a spatial resolution of 1 m [32]. The data were provided as uncompressed digital orthophoto quarter quadrangles (DOQQs) in a .tiff format. The study area is covered by 108 individual DOQQ NAIP images, representing 260,975 ha or 4.2% of the total area of the state of West Virginia.

The LIDAR data were acquired using an Optech ALTM-3100C sensor through a series of aerial flights between 28 March and 28 April 2011. The LIDAR scanner had a 36° field of view and a frequency of 70,000 Hz. The LIDAR data were acquired at a flying height of 1,524 m above the ground with an average flight speed of 250 km/h. The flight lines of the LIDAR dataset had an average of 30% overlap. The LIDAR data include elevation, intensity, up to four returns, and a vendor-provided basic classification of the points [33]. The LIDAR data were formatted as a .las version 1.2 point cloud. In total, 1164 LIDAR tiles containing a combined total of 5.6 x 109 points were used in the analysis. A preliminary investigation indicated that little change occurred during the approximately three to four-month temporal gap between the LIDAR and NAIP acquisitions.

Four land-cover classes were mapped: forest, grassland, water, and other. The forest class is primarily closed-canopy deciduous and mixed forests. The grassland class comprises areas dominated by non-woody vegetation. The water class includes both impoundments and natural waterbodies. The other class encompasses areas characterized by bare soil, exposed rock, impervious surfaces, and croplands.

2.2. Experimental Design

Sample selection includes three components: sample size, sampling region, and sampling method. The sample size specifies the number of training samples in the training set. The sampling region indicates whether samples are collected from the entire study area or only a limited sub region. The sampling method specifies the protocol for selecting samples, for example, random or deliberative.

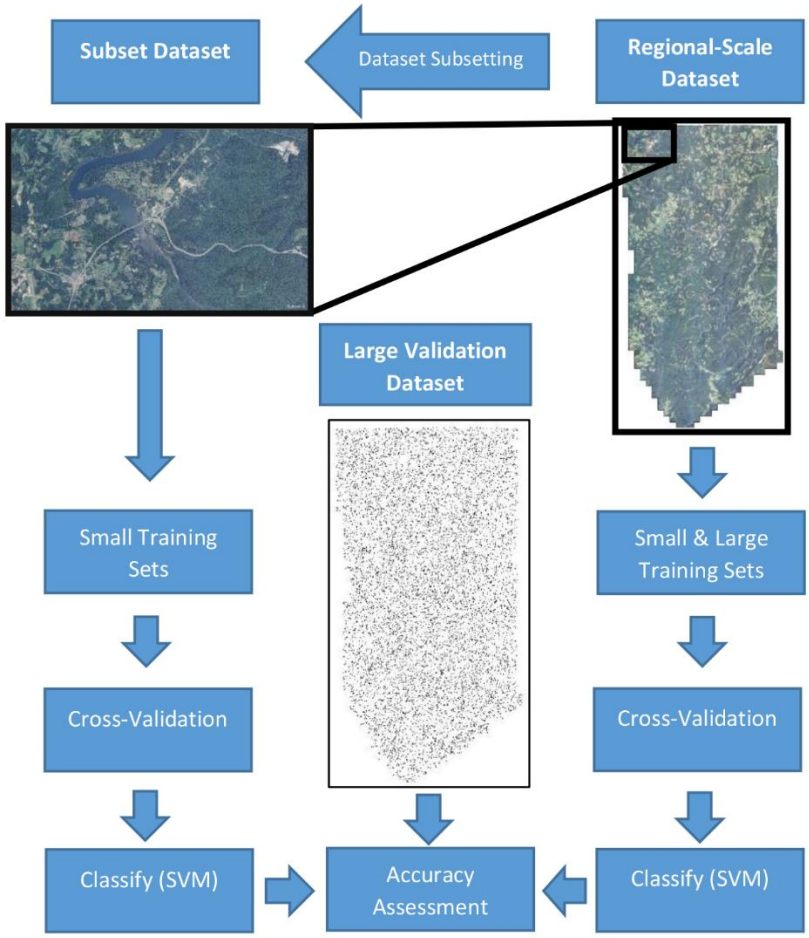

Figure 2 - Overview of the experiment workflow

In this study, four sampling methods are used to generate training data sets, which are then used with three cross-validation methods in SVM classifications (Figure 2). The samples are selected from the entire study area or from only a small geographic subset of the study area, and in all cases, the classifier is applied to the entire regional-scale dataset. The error for all classifications is evaluated using a large independent validation dataset acquired from the entire regional-scale dataset.

2.3. Data Processing

A normalized-digital surface model (nDSM) and intensity rasters were generated as input variables for the classification from the LIDAR point cloud data. The LIDAR intensity raster was generated using the first returns only and the LAS Dataset to Raster function in ArcMap 10.5.1 [34]. The LIDAR intensity data has proven to be beneficial for separating land-cover surfaces such as grassland, trees, buildings, and asphalt roads. [35–37]. The LIDAR intensity data was not normalized due to the limited LIDAR metadata which prevented the normalization for distance. Previous research indicates that LIDAR intensity information is still useful for land-cover classifications without this correction [38]. The nDSM was generated by subtracting a rasterized bare earth digital elevation model (DEM) from a digital surface model (DSM) produced from the ground and first returns, respectively. The LIDAR-derived surfaces were rasterized at 1 m, matching the pixel size of the NAIP orthoimagery. nDSMs have been demonstrated to be useful for characterizing the varying heights of natural and man-made objects in GEOBIA studies [39].

The 108 NAIP orthoimages were mosaicked into a single large NAIP image mosaic using the Mosaic Pro tool within ERDAS Imagine 2016. Color-balancing was used to reduce the radiometric variations between the NAIP images, since they were acquired from different flights and times [40]. The NAIP mosaic was clipped to the extent of the LIDAR rasters. The NAIP and LIDAR rasters were then combined to form a single layer stack with six bands: four NAIP (Red, Green, Blue, and NIR) and two LIDAR (nDSM and Intensity).

2.4. Image Segmentation

The Trimble eCognition Developer 9.3 multi-resolution segmentation (MRS) algorithm was used as the segmentation method [41]. MRS is a bottom-up region-growing segmentation approach. Equal weighting was given to all six input bands for the segmentation. Preliminary segmentation trials found that a large number of artefacts were created by the image segmentation, apparently due to the

"sawtooth" scanning pattern of the OPTECH ALTM 3100 sensor and the 1 m rasterization process [42]. A 5 x 5 pixel median filter was therefore applied to both the nDSM and Intensity rasters prior to segmentation to reduce the problem.

MRS has three user-set parameters: scale, shape, and compactness [43]. The scale parameter (SP) is regarded as the most important of the three parameters as it controls the size of the image objects [44–46]. The Estimation of Scale Parameter (ESP2) tool developed by Drăguţ et al. [45] was used to estimate the optimal scale parameter for the segmentation. The ESP2 tool generates image-objects using incrementally increasing SP values and calculates the local variance (LV) for each scale. The rate of change of the local variance (ROC-LV) is then plotted against the SP. In theory, peaks in the ROC-LV curve indicate segmentation levels in which segments most accurately delineate real world objects and thus optimal SPs for the segmentation [45].

Due to the high processing and memory demands of the ESP2 tool, three randomly selected subset areas were chosen to apply the ESP2 process rather than attempting to run the tool across the entire regional-scale dataset. The three subset tests indicated optimal SP values of 97, 97, and 104. The intermediate value of 100 was therefore chosen for the segmentation of the entire image. Alterations of the shape and compactness parameters from their defaults of 0.1 and 0.5 respectively did not seem to improve the quality of the segmentation, and therefore these values were left unchanged. The segmentation generated 474,614 image segments.

2.5. Dataset Subsetting

The subsetting tool in eCognition was used to extract the subset dataset from the regional dataset. The location of the subset was selected so that it included all four classes of interest. The total area of the subset dataset was approximately 4.19% of the area of the regional-scale dataset and comprised 21,777 image objects.

2.6. Segment Attributes Used for Classification

A total of 35 spectral and geometric attributes (Table 1) were generated for each image object (segment); these attributes were used as the predictor variables for the classification. Examples of the spectral attributes include the object's means and standard deviations for each band and the geometric attributes include object asymmetry, compactness, and roundness. The object's mean normalized difference vegetation index (NDVI) was also included as it is a commonly used spectral index used with NAIP data [47].

Table 1 - Spectral and geometric attributes of the segments

Attribute Type	Attributes	Number of Attributes
Spectral	Mean (Blue, Green, Intensity, NIR, Red, nDSM), Mode (Blue, Green, Intensity, NIR, Red, nDSM), Standard deviation (Blue, Green, Intensity, NIR, Red, nDSM), Skewness (Blue, Green, Intensity, NIR, Red, nDSM), Brightness	25
Geometric	Density, Roundness, Border length, Shape index, Area, Compactness, Volume, Rectangular fit, Asymmetry	9
Spectral Indices	Mean NDVI	1

2.7. Sample Data Selection

As image-objects are the base unit of analysis in this study, an object-based sampling approach was used for the collection of the samples. Two spatial scales were employed: a small subset and a regional scale, encompassing the entire study area. A large regional sample (n = 10,000) from the regional-scale dataset was collected to provide a benchmark representing an assumed maximum accuracy possible with this dataset. Since the subset area is 4.19% of the regional scale data set, the sample size for the subset area was set to n = 419 samples (4.19% of 10,000). This sample set is termed the small subset dataset. In addition, a small regional sample (n = 419) was selected from the entire regional scale data set to provide a direct comparison with the small subset sample dataset. In summary, three categories of

datasets were collected at two spatial scales and two sample sizes: samples from a small limited region within the study area (small subset sample) (Figure 3) and two sets of samples collected from across the study area, one encompassing a small number of samples (small regional sample) (Figure 4) and another encompassing a large number (large regional sample) (Figure 5).

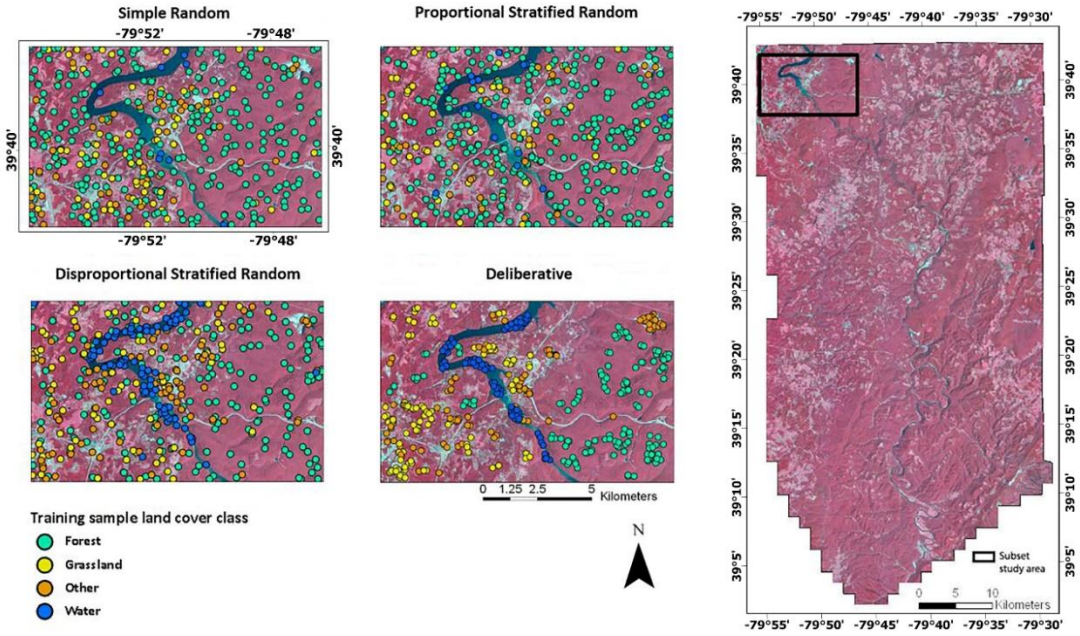

Figure 3 - The subset area location and subset training samples overlaid on false color infrared composite of National Agricultural Imagery Program (NAIP) orthoimagery (Bands 4, 1, and 2 as RGB).

For each of these three categories of spatial scales and sampling sizes, four sampling methods were employed: simple random, proportional stratified random, disproportional stratified random, and deliberative. All samples were manually labeled by the analyst. In total, 53,352 samples were collected for this analysis. The number of samples for each training and validation sample sets is summarized in Tables 2a–c.

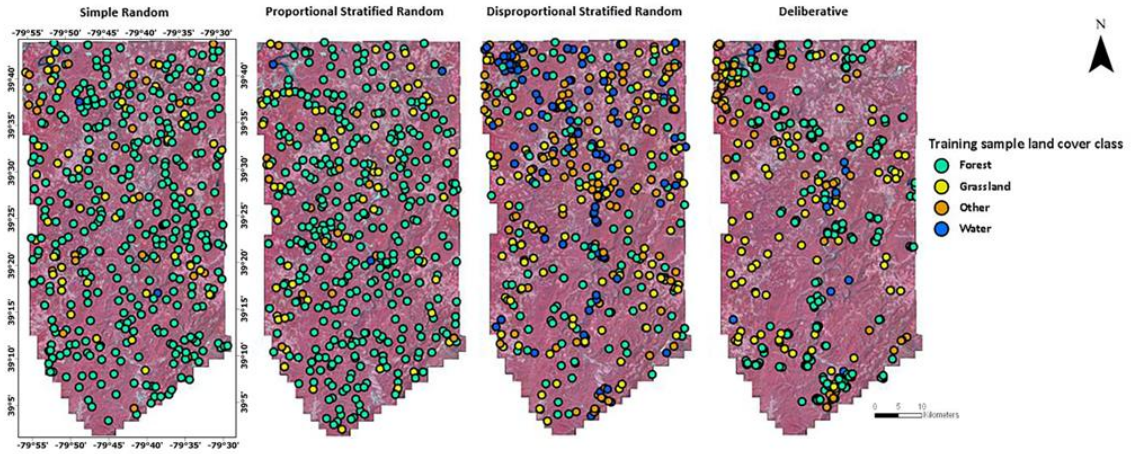

Figure 4 - The small regional training samples displayed over false color infrared composite of NAIP orthoimagery (Bands 4, 1, and 2 as RGB).

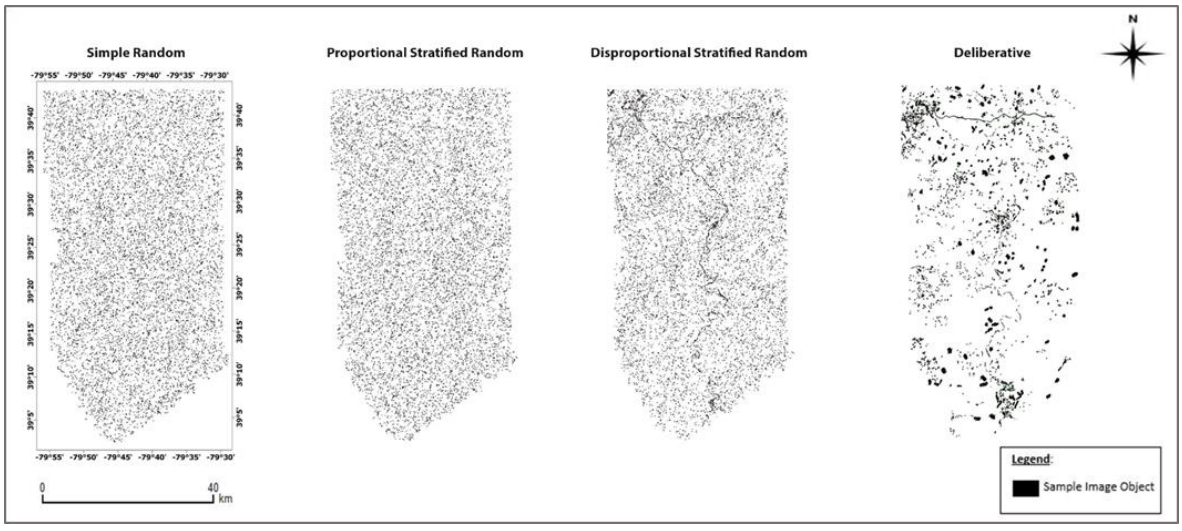

Figure 5 - Large regional-scale training sample datasets.

Table 2a- Small subset sample sets.

	Number of samples per class				
Sample Name	Forest	Grass	Other	Water	Total # of Samples
Small Subset Simple Random	290	67	53	9	419
Small Subset Proportional Stratified Random	305	59	35	20	419
Small Subset Disproportional Stratified Random	209	84	84	42	419
Small Subset Deliberative	139	100	100	80	419

Table 2b. Small regional sample sets.

	Number of samples per class				
Sample Name	Forest	Grass	Other	Water	Total # of Samples
Small Regional Simple Random	341	50	26	2	419
Small Regional Proportional Stratified Random	333	65	18	3	419
Small Regional Disproportional Stratified Random	209	84	84	42	419
Small Regional Deliberative	254	80	69	16	419

Table 2c. Small regional sample sets.

	Number of samples per class				
Sample Name	Forest	Grass	Other	Water	Total # of Samples
Large Regional Simple Random	8183	1178	600	39	10000
Large Regional Proportional Stratified Random	7984	1553	408	55	10000
Large Regional Disproportional Stratified Random	5000	2000	2000	1000	10000
Large Regional Deliberative	6087	1897	1651	365	10000

2.7.1. Simple Random Sampling

The eCognition version 9.3 client does not offer a tool for selecting random samples, and therefore the select random polygon tool in QGIS was used.

2.7.2. Proportional Stratified Random Sampling

Because a stratified approach requires a priori strata, a rule-based classification developed through the expert system was applied to both the regional-scale dataset and the subset dataset (Figure 6) to estimate the strata sizes for the subset and regional datasets.

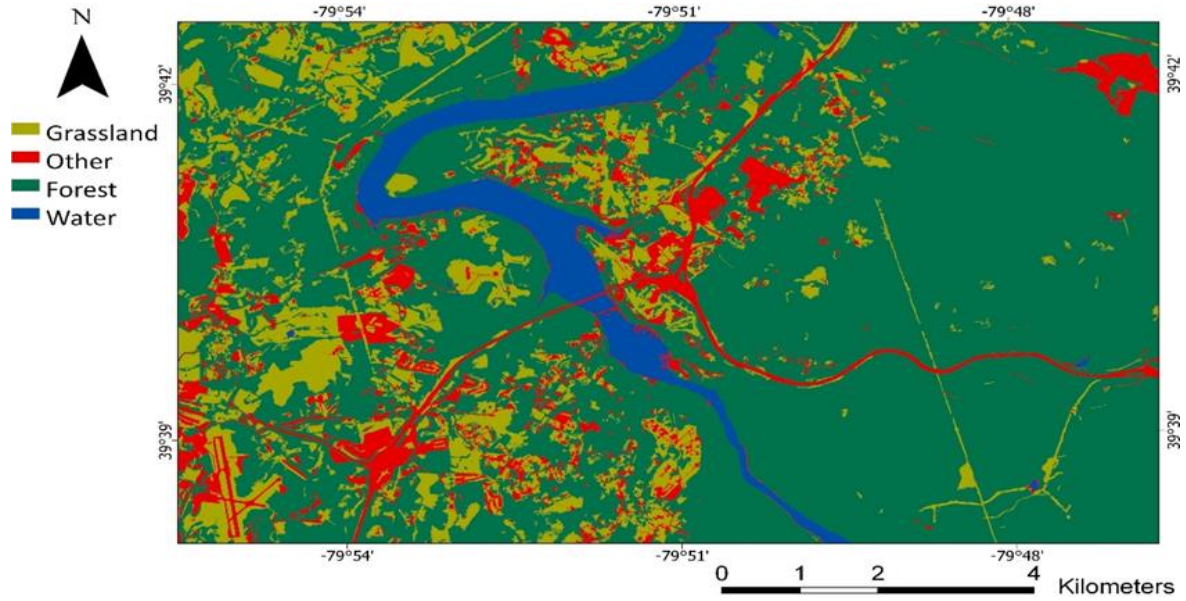

Figure 6 - Rule-based classification of subset area

The ruleset contained 16 individual rules. The accuracy of the rule-based classifications was evaluated using the samples from the large regional-scale validation dataset and had an overall accuracy of 98.1%. The strata size for both the subset and regional-scale datasets were determined by the total area occupied by each class. Table 3 summarizes the proportions of the strata for both datasets. Simple random sampling was used within each stratum to obtain samples for both the subset and regional-scale datasets.

Table 3. Class strata sizes for subset and regional datasets.

Class	Proportion of Total Area Occupied	
	Subset Dataset	Regional Dataset
Forest	72.73%	79.84%
Grassland	14.10%	15.53%
Other	8.34%	4.08%
Water	4.84%	0.55%

2.7.3. Disproportional Stratified Random Sampling

It was not possible to test an equalized stratified sampling approach because the water class is too rare to provide sufficient samples for a 25% proportion. Consequently, a disproportional stratified approach was chosen. For this sample, the class proportions were defined as 50% forest, 20% grassland, 20% other, and 10% water. These proportions were selected as intermediate values between the random and equalized stratified proportions to ensure a larger representation of the less common classes than in the random dataset. The same values were used for the small subset and small and large regional sample sets.

2.7.4. Deliberative Sampling

The deliberative sample was produced via on-screen digitizing by the analyst using the sample selection tool in eCognition Developer. No attempt was made to avoid spatial autocorrelation in the samples selected, for example, by avoiding samples that were spatially adjacent, because manual selection of samples is generally characterized by autocorrelation [48].

2.8. Cross-Validation Strategies

The cross-validation tuning methods were conducted using the trainControl function within the caret package [49] in R Studio 1.1.383. A separate classification was conducted for each cross-validation tuning method used and each sample set. The three cross-validation strategies tested were k-fold, leave-one-out, and Monte Carlo.

2.9. Supervised Classification

A radial basis function kernel (RBF) Support Vector Machines (SVM) was chosen as the supervised machine learning classifier for several reasons:

1. SVM is a commonly used supervised classifier in remote sensing analyses [17].
2. SVM is a non-parametric classifier, meaning it makes no assumption regarding the underlying data distribution. This may be advantageous for a small sample set [50].
3. SVMs are able to perform well with relatively small training datasets when compared to other commonly used classifiers.
4. SVMs are attractive for their ability to find a balance between accuracy and generalization [51].

A total of 36 individual classifications were conducted, each using a different combination of sample and validation methods: 3 categories of approaches at different spatial scales and sample sizes (small subset, small regional, and large regional) x 4 sample selection methods (simple random, proportional stratified random, disproportional stratified random, and deliberative) x 3 cross-validation tuning methods (k-fold, leave-one-out, and Monte Carlo) = 36 classifications.

Table 4 details all subset-trained and regional-trained classifications. The SVM classifications were conducted within R Studio client version 1.1.383 using the e1071 [52] and caret packages [49] on a Dell Optiplex 980 workstation with an Intel i7 2.80 GHz processor with 16.0 GB of memory running Windows 8.1 Enterprise. The processing time for all classifications were recorded using the microbenchmark package [53]. Processing runtime values should be interpreted as indications of relative speed and not as absolute values as they are highly dependent on the system architecture, CPU allocation, memory availability, and background system processes, among other factors.

Table 4. Classifications and associated abbreviations based on the sample selection method, training sample size, region of area collected, and cross-validation method.

		Cross-Validation Method								
		k-Fold (KF)	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**	**k-Fold (KF)**	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**	**k-Fold (KF)**	**Monte Carlo (MC)**	**Leave-One-Out (LOO)**
		Small Subset-Trained Classification			**Small Regional-Scale Trained Classification**			**Large Regional-Scale Trained Classification**		
Sample Selection Method	**Simple Random (SR)**	Small-Subset-(SR-KF)	Small-Subset-(SR-MC)	Small-Subset-(SR-LOO)	Small-Regional-(SR-KF)	Small-Regional-(SR-MC)	Small-Regional-(SR-LOO)	Large-Regional-(SR-KF)	Large-Regional-(SR-MC)	Large-Regional-(SR-LOO)
	Proportional Stratified Random (PSTR)	Small-Subset-(PSTR-KF)	Small-Subset-(PSTR-MC)	Small-Subset-(PSTR-LOO)	Small-Regional-(PSTR-KF)	Small-Regional-(PSTR-MC)	Small-Regional-(PSTR-LOO)	Large-Regional-(PSTR-KF)	Large-Regional-(PSTR-MC)	Large-Regional-(PSTR-LOO)
	Disproportional Stratified Random (DSTR)	Small-Subset-(DSTR-KF)	Small-Subset-(DSTR-MC)	Small-Subset-(DSTR-LOO)	Small-Regional-(DSTR-KF)	Small-Regional-(DSTR-MC)	Small-Regional-(DSTR-LOO)	Large-Regional-(DSTR-KF)	Large-Regional-(DSTR-MC)	Large-Regional-(DSTR-LOO)
	Deliberative (DL)	Small-Subset-(DL-KF)	Small-Subset-(DL-MC)	Small-Subset-(DL-LOO)	Small-Regional-(DL-KF)	Small-Regional-(DL-MC)	Small-Regional-(DL-LOO)	Large-Regional-(DL-KF)	Large-Regional-(DL-MC)	Large-Regional-(DL-LOO)

2.10. Error Assessment

Each of the trained classifications was tested against a large, randomly sampled validation dataset (n = 10,000) collected from the regional-scale dataset. Results for each classification were reported in a confusion matrix programmed via the caret package in the R statistical client. User's and producer's accuracies were calculated as well as overall accuracy and the kappa coefficient. Additionally, McNemar's test [54] was used to evaluate the statistical significance of differences observed between the k-fold tuned classifications. McNemar's test is a non-parametric evaluation of the statistical differences between two classifications with related samples [55]. A p-value smaller than 0.05 indicates a one-sided 95% confidence that the differences in accuracy between the classifications are statistically significant.

3. Results and Discussion

3.1. Performance of Sample Selection Methods

Figure 7 summarizes the overall accuracies of the classification of the entire regional dataset using the various training samples and k-fold (k = 10) cross-validation. Within each spatial scale and sample size (i.e., subset, small regional, and large regional), the disproportional stratified random (DSTR) samples consistently resulted in the highest overall accuracy although it is notable that variations between the performance of the classifications trained using the different statistical-based sampling methods were small, less than 2%.

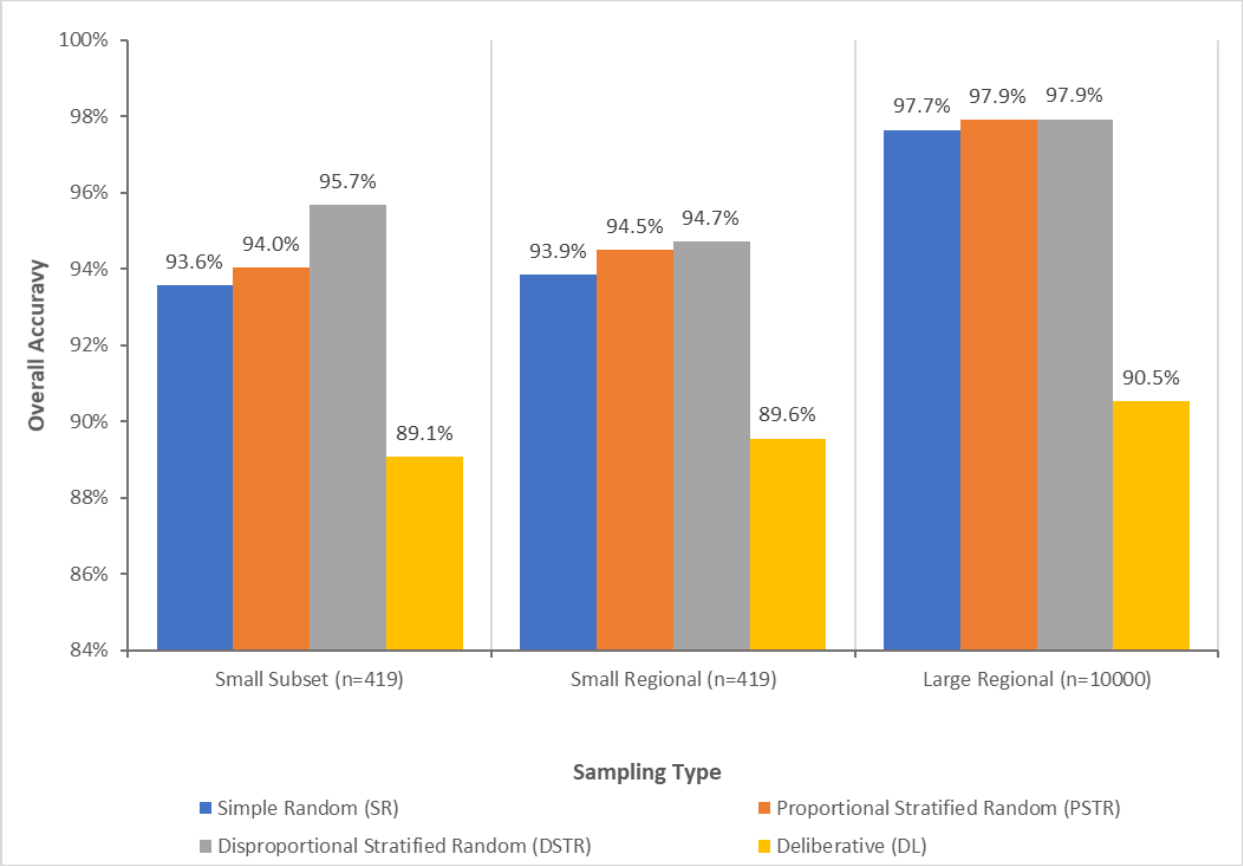

Figure 7 - Overall accuracies of the regional classifications using small subset, small regional, and large regional training datasets and k-fold (k=10) cross-validation tuning.

Despite the small differences between some of the classification accuracies, the McNemar's test results, shown in Table 5, indicate that most of the differences were statistically significant. The only

exceptions were the differences in the classification accuracies for the large regional-trained SR, PSTR, and DSTR, which indicates that when the sample size is very large (*n* = 10,000 in this case), differences between sampling methods is less important.

Classifications trained with the SR sample resulted in a slightly lower accuracy than those trained with the PSTR and DSTR samples. This suggests that sample stratification is advantageous for SVM classifiers, as stratification ensures that rare classes are sampled at a rate that is either equal to or greater than their proportion in the dataset, depending on whether proportional or disproportional stratified approaches are selected.

The DSTR sampling method was designed to provide a much larger number of samples from minority classes, such as the water and other classes, than the simple random or proportional stratified random sampling methods (Table 2a–c). Using SR sampling, the number of samples from the minority classes may vary greatly, depending on random chance, especially when the total number of samples is small (e.g., 419 in this case). This can be seen in the small-regional SR and PSTR sample sets, where only 2 and 3 samples, respectively, were collected for water (Table 2b). The number of samples selected for the rare classes is important; Waske et al. [56] found that SVM was negatively affected by dataset imbalance. Stehman [57] also found that sample stratification resulted in improved classification accuracy due to an increased sample selection from the minority classes. The results of our study emphasize the value of disproportional stratified sample selection to reduce class imbalance and ensure minority class representation in the training set.

*Table 5. McNemar's test p-values for small subset, small regional, and large regional-trained classifications using k-fold (k=10) cross-validation tuning. (*Indicates the differences between classifications that are statistically significant, $p < 0.05$.)*

Subset-SR-KF	Subset-PSTR-KF	Subset-DSTR-KF	Subset-DL-KF	Small-Regional-SR-KF	Small-Regional-PSTR-KF	Small-Regional-DSTR-KF	Small-Regional-DL-KF	Large-Regional-SR-KF	Large-Regional-PSTR-KF	Large-Regional-DSTR-KF	Large-Regional-DL-KF	
	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Subset-SR-KF
		0.007*	<0.001*	<0.001*	<0.001*	0.004*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Subset-PSTR-KF
			<0.001*	0.043*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Subset-DSTR-KF
				<0.001*	<0.001*	<0.001*	0.009*	<0.001*	<0.001*	<0.001*	<0.001*	Subset-DL-KF
					0.013*	0.002*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Small-Regional-SR-KF
						0.031*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Small-Regional-PSTR-KF
							<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	Small-Regional-DSTR-KF
								<0.001*	<0.001*	<0.001*	<0.001*	Small-Regional-DL-KF
									0.108	0.113	<0.001*	Large-Regional-SR-KF
										0.162	<0.001*	Large-Regional-PSTR-KF
											<0.001*	Large-Regional-DSTR-KF
												Large-Regional-DL-KF

The classifications trained from the deliberative (DL) samples consistently had lower accuracies across all sample sets (Figure 7). The low accuracy for the classifications with the DL samples indicates that samples acquired though expert selection of the training data did not adequately characterize the dataset. Notably, the DL samples have higher spatial autocorrelation than samples collected via the statistical-based methods (Figure 5). This is not surprising; as noted previously, human-based

deliberative sampling has a high potential for spatial autocorrelation [48]. High spatial autocorrelation in sample sets may result in a reduction in the effective sample size [13]. A univariate Moran's I test indicates that the small subset, small regional, and large regional DL samples all show positive spatial autocorrelation, with values of 0.950, 0.692, and 0.985, respectively. While the SR, PSTR, and DSTR samples also contained positive spatial autocorrelation, ranging from 0.208 (subset-SR) to 0.661 (large-regional-DSTR), they showed less positive spatial autocorrelation than all DL sample sets. Stratified sampling, especially disproportional stratified sampling, tends to favor some autocorrelation, since samples are, by definition, not completely random.

The similar performance between the classifications trained from the small subset and small regional-scale samples and the much higher accuracy reported from the large regional-scale sample indicate that the geographic location of the sample may not be as important as the sample size in determining the accuracy of the supervised SVM classifications. This is notable as selecting a small sample from a subset area is less expensive in terms of effort than selecting a small sample from a regional-scale area, especially if field data collection is needed. It should also be mentioned that the regional-scale area in this analysis was generally homogenous, which allowed the selection of a single subset area that contained many examples of all four classes of interest. In areas or datasets that are more heterogeneous or contain extreme minority classes limited to separate geographic regions of a regional-scale dataset, multiple subset areas may be needed for subset-based sampling to be effective. The fact that the classifications trained from both small sample sets were less accurate than the large regional-trained benchmark classification emphasizes that large numbers of statistical-based samples can raise the accuracy of SVM classifications substantially.

3.2. Performance of Cross-Validation Tuning Methods

There was no consistent pattern for the overall accuracy of the classifications trained from the small subset samples and tuned using the three cross-validation methods (k-fold, Monte Carlo, and

leave-one-out). As seen in Figure 8a, when the SR samples were used as training data, the k-fold (KF) method provided slightly higher accuracy than the Monte Carlo (MC) or leave-one-out (LOO) cross-validation. MC proved the best method for the PSTR samples. KF, MC, and LOO tuning resulted in equal overall accuracies for the DSTR samples. Overall, the differences in overall accuracy between the tuning methods on the small subset statistical-sample trained classifications was less than 1%.

The cross-validation tuning methods using the small regional SR, PSTR, and DSTR training data applied to the SVM classification all showed high values for overall accuracy (Figure 8b) and inconsistent results for the different tuning methods, similar to the results of the small subset training data. LOO had slightly lower performance on the SR classifications, but this decrease in performance was also less than 1%. The DSTR classifications had the highest overall accuracy, irrespective of the cross-validation tuning methods, with the MC- and LOO-tuned DSTR classifications resulting in 94.8% and the k-fold tuned-DSTR resulting in 94.7% overall accuracy.

For the large-regional statistical sample-trained classifications, the MC- and KF-tuned classifications consistently outperformed LOO. This indicates that LOO is less effective for tuning than KF and MC when dealing with large statistical-based sample sets. Both MC and KF had the same overall accuracy for the large-regional SR, PSTR, and DSTR classifications.

The deliberative-trained classifications for both the small subset and regional scales had much lower performances than the statistical based classifications. For the small-regional DL classifications, LOO matched the overall accuracy of KF and MC at 89.6% (Figure 8b). However, for both the small-subset (Figure 8a) and large-regional DL classifications (Figure 8c), LOO tuning improved overall accuracy by 3.6% and 1.2%, respectively over the KF-DL classifications.

The confusion matrices of the small subset-DL-KF (Table 6), MC (Table 7), and LOO (Table 8) show that the increase in performance of the small subset-DL-LOO classification was due to a substantial

increase in the producer's accuracy of the forest class and the increased user's accuracy of the grassland class, though at the cost of decreases in the other class accuracies, most notably in the grassland producer's accuracy. As the forest and grassland classes combined make up 93.4% of the validation dataset, improving the average class accuracies of these two classes led to a marked improvement of the overall accuracy.

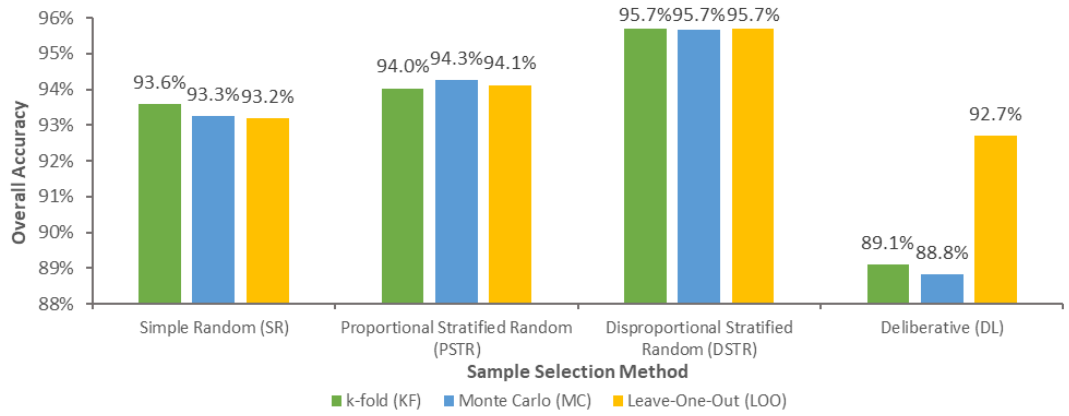

Figure 8a. Overall accuracies of the small-subset SVM training data classifications using k-fold (k=10), Monte Carlo, and leave-one-out cross-validation tuning.

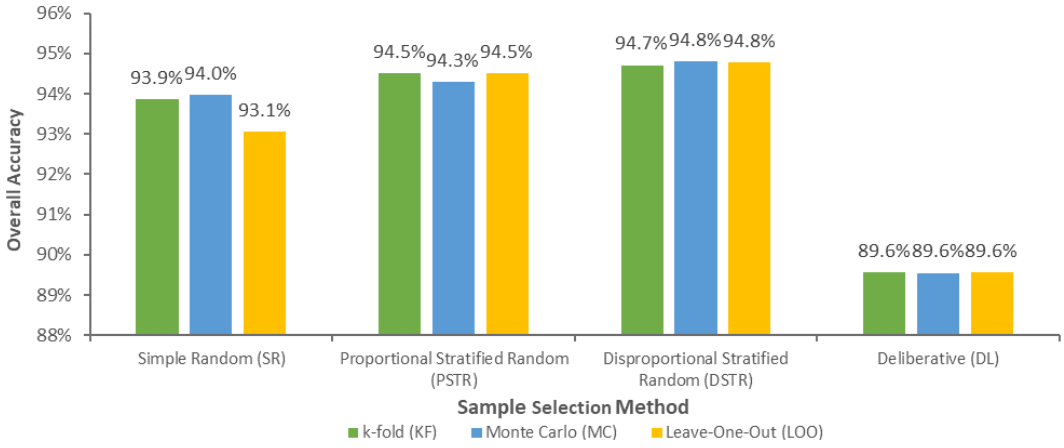

Figure 8b. Overall accuracies of the small-regional training data SVM classifications using k-fold (k=10), Monte Carlo, and leave-one-out cross-validation tuning.

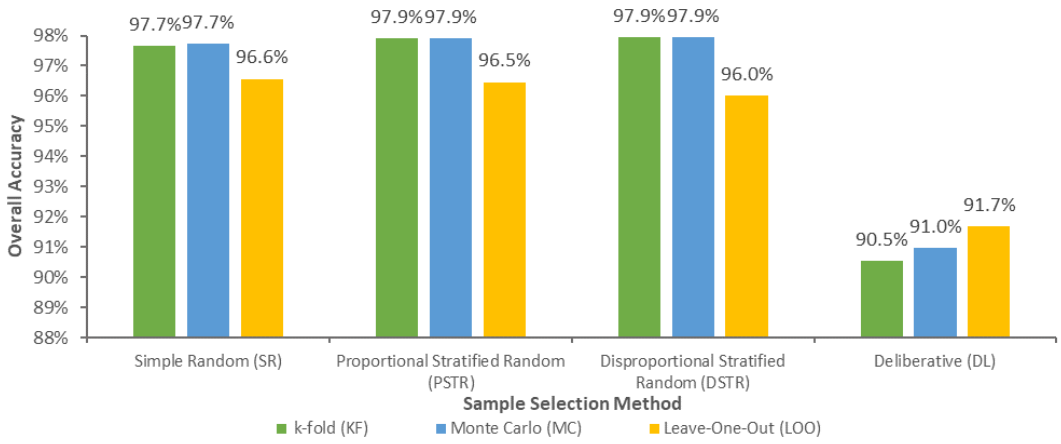

Figure 8c. Overall accuracies of the classifications using large-regional training data and k-fold (k=10), Monte Carlo, and leave-one-out cross-validation tuning.

Table 6. Confusion matrix for the classification trained with the subset-DL-KF data set.

		Reference data (No. objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's accuracy**
Classified data (No. objects)	**Forest**	7238	31	4	0	7273	99.5%
	Grassland	699	1151	85	0	1935	59.5%
	Other	136	73	479	28	716	66.9%
	Water	12	1	22	41	76	53.9%
	Total	8085	1256	590	69	10000	**Overall accuracy: 89.1%**
	Producer's accuracy	89.5%	91.6%	81.2%	59.4%		

Table 7. Confusion Matrix for the classification trained with the subset-DL-MC data set.

		Reference data (No. objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's accuracy**
Classified data (No. objects)	**Forest**	7723	36	6	0	7265	99.4%
	Grassland	723	1133	80	0	1936	58.5%
	Other	129	86	487	29	731	66.6%
	Water	10	1	17	40	68	58.8%
	Total	8085	1256	590	69	10000	**Overall accuracy: 88.8%**
	Producer's accuracy	89.3%	90.2%	82.5%	58.0%		

Table 8. Confusion Matrix for the classification trained with the subset-DL-LOO data set.

		Reference data (No. objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's accuracy**
Classified data (No. objects)	**Forest**	7728	174	5	0	7907	97.7%
	Grassland	176	1034	95	1	1306	79.2%
	Other	176	48	466	26	716	65.1%
	Water	5	0	24	42	71	59.2%
	Total	8085	1256	590	69	10000	**Overall accuracy: 92.7%**
	Producer's accuracy	95.6%	82.3%	79.0%	60.9%		

However, both the leave-one-out and Monte Carlo tuning required longer processing time than the k-fold (k = 10) cross-validation tuning (Table 9). When sample sizes become very large (n = 10,000), leave-one-out tuning may become prohibitively slow; though with advances in processor technology, this may be less of a concern for the future.

Table 9. Processing time metrics.

Classification	Processing time (seconds)
SVM-Subset-KF	10
SVM-Large-Regional-KF	468
SVM-Subset-MC	17
SVM-Large-Regional-MC	876
SVM-Subset-LOO	313
SVM-Large-Regional-LOO	489,960

Since no cross-validation method was consistently superior for tuning SVM classifications trained from the SR, PSTR, and DSTR sample sets and for all sample sets, k-fold may be the most effective and efficient method for cross-validation parameter tuning for SVM classifiers.

4. Conclusions

This investigation explored the effects sample acquisition method, sample geographic distribution, and cross-validation tuning methods in regional-scale land-cover classifications of HR remotely sensed data. Based upon the results presented in this analysis, a random sample, possibly

combined with stratification techniques to ensure adequate representation of minority classes within training sample sets, is recommended. Deliberative samples should be avoided, possibly because of the tendency of humans to collect excessively highly autocorrelated samples. Stehman, [57] recommends using an underlying systematic sampling scheme to minimize spatial autocorrelation in sample sets.

Classifications trained from the small subset-based sample sets were found to have comparable performance to classifications trained from small sample sets acquired in a dispersed manner across the entire regional-scale study site. This is an important finding because if sample selection is expensive, especially if field checking is required, a relatively small sample set collected from a subset area of the regional-scale study area can be used. However, it is important to note that since the study area for this analysis was broadly homogenous, it was possible to select a single subset area that contained adequate examples of all four classes of interest for training data collection. In more heterogeneous environments, multiple subset areas may be needed to obtain the samples. Future research is needed on large-scale sample selection strategies in highly heterogeneous environments.

The relative accuracy of classifications produced with k-fold (k = 10), leave-one-out, and Monte Carlo cross-validation tuning methods when trained with the small subset, small regional, and large regional SR, PSTR, and DSTR data sets were not consistent. As the Monte Carlo and leave-one-out cross-validation tuning methods required greater processing resources and time, the k-fold cross-validation method may be preferable, especially for large sample sets. Regarding deliberative sampling methods, in both the small subset and large regional classifications, leave-one-out cross-validation tuning was more effective in increasing classifier performance when compared to the k-fold and Monte Carlo tuning.

In summary, for large regional-scale HR classifications, deliberative sampling should be avoided not only for accuracy assessment data but also for training data collection. Random samples are

preferable, and data collected randomly from a small subset region is adequate, at least in relatively homogenous areas. Disproportional stratified sampling can be used to reduce the effect of imbalanced samples. Tuning is important, though the type of method used does not seem to have a large effect. k-fold tuning is possibly a good choice because it is relatively rapid.

References

1. Fassnacht, F.E.; Hartig, F.; Latifi, H.; Berger, C.; Hernandez, J.; Corvalan, P.; Koch, B. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens. Environ.* **2014**, *154*, 102–114, doi:10.1016/j.rse.2014.07.028.
2. Guo, Y.; Ma, L.; Zhu, F.; Liu, F. Selecting Training Samples from Large-Scale Remote-Sensing Samples Using an Active Learning Algorithm. In *Computational Intelligence and Intelligent Systems*; Li, K., Li, J., Liu, Y., Castiglione, A., Eds.; Springer: Singapore, 2016; pp. 40–51, doi:10.1007/978-981-10-0356-1_5.
3. Lu, D.; Weng, Q. A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* **2007**, *28*, 823–870, doi:10.1080/01431160600746456.
4. Foody, G.M. Sample size determination for image classification accuracy assessment and comparison. *Int. J. Remote Sens.* **2009**, *30*, 5273–5291, doi:10.1080/01431160903130937.
5. Jin, H.; Stehman, S.V.; Mountrakis, G. Assessing the impact of training sample selection of accuracy of an urban classification: A case study in Denver, Colorado. *Int. J. Remote Sens.* **2014**, *35*, 2067–2081, doi:10.1080/01431161.2014.885152.
6. Radoux, J.; Bogaert, P.; Fasbender, D.; Defourny, P. Thematic accuracy assessment of geographic object-based image classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 895–911, doi:10.1080/13658816.2010.498378.
7. Stehman, S.V. Impact of sample size allocation when using stratified random sampling to estimate accuracy and area of land-cover change. *Remote Sens. Lett.* **2012**, *3*, 111–120, doi:10.1080/01431161.2010.541950.
8. Ma, L.; Li, M.; Ma, X.; Cheng, K.; Du, P.; Liu, Y. A review of supervised object-based land-cover image classification. *ISPRS J. Photogramm. Remote Sens*. **2017**, *130*, 277–293, doi:10.1016/j.isprsjprs.2017.06.001.

9. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16, doi:10.1016/j.isprsjprs.2009.06.004.
10. Foody, G.M.; Pal, M.; Rocchini, D.; Garzon-Lopez, C.X.; Bastin, L. The Sensitivity of mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 1–20.
11. Congalton, R.G. A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data. *Remote Sens. Environ.* **1991**, *37*, 35–46.
12. Foody, G.M.; Mathur, A.; Sanchez-Hernandez, C.; Boyd, D.S. Training Set Size Requirements for the Classification of a Specific Class. *Remote Sens. Environ.* **2006**, *104*, 1–14.
13. Mu, X.; Hu, M.; Song, W.; Ruan, G.; Ge, Y.; Wang, J.; Huang, S.; Yan, G. Evaluation of Sampling Methods for Validation of Remotely Sensed Fractional Vegetation Cover. *Remote Sens.* **2015**, *7*, 16164–16182, doi:10.3390/rs71215817.
14. Chen, D.M.; Stow, D. The Effect of Training Strategies on Supervised Classification at Different Spatial Resolutions. *Photogramm. Eng. Remote Sens.* **2002**, *68*, 1155–1161.
15. Chen, D.; Stow, D.A.; Gong, P. Examining the effect of spatial resolution and texture windows size on classification accuracy: An urban environment case. *Int. J. Remote Sens.* **2004**, *25*, 2177–2192.
16. Congalton, R.G. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. *Photogramm. Eng. Remote Sens.* **1988**, *54*, 593–600.
17. Maxwell, A.E.; Warner, T.A.; Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **2018**, *39*, 2784–2817, doi:10.1080/01431161.2018.1433343.

18. Stehman, S.V. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. *Int. J. Remote Sens.* **2014**, *35*, 4923–4939, doi:10.1080/01431161.2014.930207.
19. Stehman, S.V.; Foody, G.M. Accuracy assessment. In *The SAGE Handbook of Remote Sensing*; Warner, T.A., Nellis, M.D., Foody, G.M., Eds.; Sage Publications Ltd.: London, UK, 2009; pp. 129–145, ISBN 9781412936163.
20. Pal, M.; Foody, G.M. Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 1344–1355, doi:10.1109/JSTARS.2012.2215310.
21. Demir, B.; Minello, L.; Bruzzone, L. An Effective Strategy to Reduce the Labeling Cost in the Definition of Training Sets by Active Learning. *IEEE Geosci. Remote Sens. Lett*. **2014**, *11*, 79–83, doi:10.1109/LGRS.2013.2246539.
22. Wuttke, S.; Middlemann, W.; Stilla, U. Concept for a compound analysis in active learning remote sensing. In Proceedings of the International Archives of the Photogrammetry, Remote Sensing, and Spatial Information Sciences, Munich, Germany, 25–27 March 2015; Volume XL-3(W2), pp. 273–279, doi:10.5194/isprsarchives-XL-3-W2-273-2015.
23. Babcock, C.; Finely, A.O.; Bradford, J.B.; Kolka, R.K.; Birdsey, R.A.; Ryan, M.G. LiDAR based prediction of forest biomass using hierarchial models with spatially varying coefficients. *Remote Sens. Environ.* **2015**, *169*, 113–127, doi:10.1016/j.rse.2015.07.028.
24. Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 5372–5375, doi:10.1109/IGARSS.2012.6352393.

25. Cracknell, M.J.; Reading, A.M. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Comput. Geosci.* **2014**, *63*, 22–33, doi:10.1016/j.cageo.2013.10.008.
26. Sharma, R.C.; Hara, K.; Hirayama, H. A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data. *Scientifica* **2017**, *2017*, 9806479, doi:10.1155/2017/9806479.
27. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272, doi:10.1016/j.rse.2011.11.020.
28. Stone, M. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc.* **1974**, *36*, 111–147.
29. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, USA, 2009.
30. Picard, R.R.; Cook, R.D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *387*, 575–583.
31. Braun, E.L. *Deciduous Forests of Eastern North America*; Hafner Publishing Company: New York, NY, USA, 1950.
32. Maxwell, A.E.; Strager, M.P.; Warner, T.A.; Zegre, N.P.; Yuill, C.B. Comparison of NAIP orthophotography and RapidEye satellite imagery for mapping of mining and mine reclamation. *GISci. Remote Sens.* **2014**, *51*, 301–320, doi:10.1080/15481603.2014.912874.
33. WVU NRAC. Aerial Lidar Acquistion Report: Preston County and North Branch (Potomac) LIDAR *.LAS 1.2 Data Comprehensive and Bare Earth. West Virginia Department of Environmental Protection. Available online:

http://wvgis.wvu.edu/lidar/data/WVDEP_2011_Deliverable4/WVDEP_deliverable_4_Project_Report.pdf (accessed on 1 December 2018).

34. ESRI. *ArcGIS Desktop: Release 10.5.1*; Environmental Systems Research Institute: Redlands, CA, USA, 2017.
35. Charaniya, A.P.; Manduchi, R.; Lodha, S.K. Supervised parametric classification of aerial LIDAR data. In Proceedings of the IEEE 2004 Conferences on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004.
36. Kashani, A.G.; Olsen, M.; Parrish, C.; Wilson, N. A Review of LIDAR Radiometric Processing: From Ad Hoc Intensity correction to Rigorous Radiometric Calibration. *Sensors* **2015**, *15*, 28099–28128, doi:10.3390/s151128099.
37. Song, J.H.; Han, S.H.; Yu, K.Y.; Kim, Y.I. Assessing the possibility of land-cover classification using LIDAR intensity data. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2002**, *34*, 259–262.
38. Maxwell, A.E.; Warner, T.A.; Strager, M.P.; Conley, J.F.; Sharp, A.L. Assessing machine learning algorithms and image- and LiDAR-derived variables for GEOBIA classification of mining and mine reclamation. *Int. J. Remote Sens.* **2015**, *36*, 954–978, doi:10.1080/01431161.2014.1001086.
39. Beşol, B.; Alganci, U.; Sertel, E. The use of object based classification with nDSM to increase the accuracy of building detection. In Proceedings of the 25th Signal Processing and Communications Applications Conference (SIU), Antalya, Turkey, 15–18 May 2017.
40. Lear, R.F. NAIP Quality Samples. United States Department of Agriculture Aerial Photography Field Office. Available Online: https://www.fsa.usda.gov/Internet/FSA_File/naip_quality_samples_pdf.pdf (accessed on 28 December 2018).
41. Trimble. *Trimble eCognition Suite 9.3.2*; Trimble Germany GmbH: Munich, Germany, 2018.

42. Petrie, G.; Toth, C.K. Airborne and Spaceborne Laser Profilers and Scanners. In *Topographic Laser Ranging and Scanning: Principles and Processing*; Shan, J., Toth, C.K., Eds.; CRC Press: Boca Raton, FL, USA, 2008.
43. Baatz, M.; Schäpe, A. Multiresolution segmentation—An optimization approach for high quality multi-scale image segmentation. In Proceedings of the Angewandte Geographische Informations-Verarbeitung XII, Wichmann Verlag, Karlsruhe, Germany, 2000; pp. 12–23.
44. Belgiu, M.; Drăgut, L. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS J. Photogramm. Remote Sens.* **2014**, *96*, 67–75, doi:10.1016/j.isprsjprs.2014.07.002.
45. Drăgut, L.; Csillik, O.; Eisank, C.; Tiede, D. Automated parameterization for multi-scale image segmentation on multiple layers. *ISPRS J. Photogramm. Remote Sens.* **2014**, *88*, 119–127, doi:10.1016/j.isprsjprs.2013.11.018.
46. Kim, M.; Warner, T.A.; Madden, M.; Atkinson, D. Multi-scale texture segmentation and classification of salt marsh using digital aerial imagery with very high spatial resolution. *Int. J. Remote Sens.* **2011**, *32*, 2825–2850.
47. Maguigan, M.; Rodgers, J.; Dash, P.; Meng, Q. Assessing Net Primary Production in Montane Wetlands from Proximal, Airborne, and Satellite Remote Sensing. *Adv. Remote Sens.* **2016**, *5*, 118–130, doi:10.4236/ars.2016.52010.
48. Griffith, D.A. Establishing Qualitative Geographic Sample Size in the Presence of Spatial Autocorrelation. *Ann. Assoc. Am. Geogr.* **2013**, *103*, 1107–1122, doi:10.1080/00045608.2013.776884.
49. Kuhn, M. Caret: Classification and Regression Training. R package Version 6.0-71. 2016. Available online: https://CRAN.R-project.org/package=caret (accessed on 21 Feb 2018).

50. Scheuenemeyer, J.H.; Drew, L.J. *Statistics for Earth and Environmental Scientists*; John Wiley & Sons: Hoboken, NJ, USA, 2010; ISBN 9780470650707.
51. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens.* **2011**, *66*, 247–259, doi:10.1016/j.isprsjprs.2010.11.001.
52. Meyer, D. Support Vector Machines: The Interface to Libsvm in Package e1071. R Package Version 6.0-71. 2012. Available online: https://CRAN.R-project.org/package=e1071 (accessed on 21 Feb 2018).
53. Ulrich, J.M. Microbenchmark: Accurate Timing Functions. R Package Version 1.4-4. 2018. Available online: https://cran.r-project.org/web/packages/microbenchmark/microbenchmark.pdf (accessed on 21 Feb 2018).
54. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157, doi:10.1007/BF02295996.
55. Foody, G.M. Thematic Map Comparison: Evaluating the Statistical Significance of Differences in Classification Accuracy. *Photogramm. Eng. Remote Sens.* **2004**, *70*, 627–633, doi:10.14358/PERS.70.5.627.
56. Waske, B.; Benediktsson, J.A.; Sveinsson, J.R. *Classifying Remote Sensing Data with Support Vector Machines and Imbalanced Training Data*; Benediktsson, J.A., Kittler, J., Roli, F., Eds.; CMS 2009, LNCS 5519; Spring: Berlin/Heidleberg, Germany, 2009; pp. 375–384.
57. Stehman, S.V. Sampling designs for accuracy assessment of land cover. *Int. J. Remote Sens.* **2009**, *30*, 5243–5272, doi:10.1080/01431160903131000.

Chapter 3

What is the Optimal Training Sample Size for Common Machine Learning Classifiers?

Abstract

The size of the training data set is a major determinant of classification accuracy. Nevertheless, the collection of a large training data set for supervised classifiers can be a challenge, especially for regional studies covering a large area, which may be typical of many real world applied projects. This work investigates how variations in training sample size, ranging from a very large sample (n = 10,000) to a very small sample (n = 40), affect the performance of five supervised machine learning algorithms applied to classify large, regional-scale high spatial resolution (HR) (1 – 5 m) remotely sensed data. The performance of five supervised machine learning algorithms is evaluated: support vector machines (SVM), random forests (RF), *k*-nearest neighbors (*k*-NN), single-layer perceptron neural networks (NEU), and learning vector quantization (LVQ). RF, the algorithm with the highest overall accuracy, was notable for its negligible decrease in overall accuracy, 0.7%, when training sample size decreased from 10,000 to 315 samples. NEU and SVM were the most sensitive to decreasing sample size, with NEU classifications having slightly higher overall accuracy than SVM classifications. NEU however required a longer processing time. The *k*-NN classifier saw less of a drop in overall accuracy than NEU and SVM as training set size decreased, however the overall accuracies of *k*-NN were typically less than RF, NEU, and SVM classifiers. LVQ had on average the lowest overall accuracy of all five methods, but was relatively insensitive to sample size, down to the smallest sample sizes. Overall, due to its relatively high accuracy with small sample sets, and minimal variations in overall accuracy between very large and small sample sets, as well as short processing time, in this case, RF is a good classifier for regional-scale land-cover classifications of HR remotely sensed data, especially when training data are scarce.

1. Introduction

One of the key determinants of classification accuracy is the training sample size (Foody et al., 2006), with larger training sets typically resulting in superior performance compared to smaller training sets. However, in applied remote sensing analyses, training data may be limited and expensive to obtain, especially if field observations are needed. In circumstances where the number of training data is limited, it would be advantageous to know the relative dependence of machine learning classifiers on sample size. For example, an analyst may want to know the potential for increased classification accuracy if additional resources were invested in increasing the number of training samples. Alternatively, if a very large sample size is available, does this potentially affect the classifier choice?

The existing literature on training sample size and its effect on classification accuracy offers only partial insight into these questions. Most previous studies comparing supervised machine learning classifier accuracy have used a single, fixed training sample size (Maxwell et al., 2018; Raczko and Zagajewski, 2017; Samaneigo and Schulz, 2009), and thus have ignored the effects of variation in sample size. Conversely, investigations that have examined the effects of sample size (Foody et al., 1995; Foody et al., 2006; Millard and Richardson, 2015) have generally focused on a single classifier, making it difficult to compare the relative dependence of machine learning classifiers on sample size.

The small number of studies that have investigated varying training set size on multiple supervised classifiers have generally considered only a narrow range in sample sizes, and often focused on other characteristics of the training set, such as class imbalance (Heydari and Mountrakis, 2018; Noi and Kappas, 2018) or feature set dimensionality (Myburgh and van Niekerk, 2014). For example, an important study by Qian et al. (2015) investigated the effects of sample size on four machine learning classifiers. However, their experiment explored only a small range of sample sizes, 25 – 200, and collected from a relatively small study area. Furthermore, although they included classification and

regression tree (CART) classification, they did not include the popular random forest classifier. Shang et al. (2018) also examined the effects of training set size on various supervised classifiers in a GEOBIA classification. However, Shang et al. (2018) examined a narrow range of sample sizes, 5-50 samples per class. They also conducted their investigation using Landsat-8 OLI imagery, a medium spatial resolution dataset, which was applied to a single district within Beijing. Furthermore, their study, like that of Qian et al. (2018), did not include either neural networks or *k*-nearest neighbors classifiers.

This paper therefore furthers the investigation into the effects of sample size on supervised classifiers by examining a broad range of training sample sizes, ranging from a very large sample size of 10,000, with each class having a minimum of 1,000 samples, to a very small training sample size of 40, where certain classes may have as few as 4 training samples. The effect of sample size is compared for five supervised machine learning classifiers, support vector machines (SVM), random forests (RF), *k*-nearest neighbors (*k*-NN), single layer perceptron neural networks (NEU), and learning vector quantization (LVQ). The accuracy of the classifications is evaluated with a large, independent validation sample set.

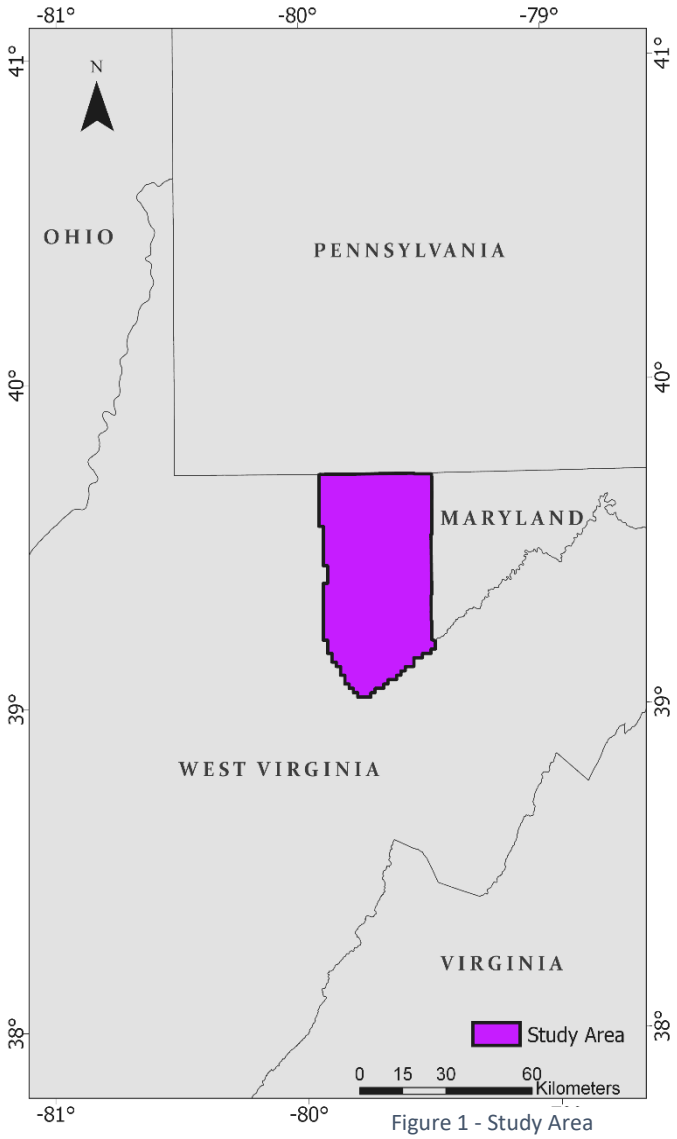

Figure 1 - Study Area

As most previous investigations comparing supervised machine learning classifiers dependence on sample size have employed relatively small test areas, this analysis examines classifier response to varying training sample sizes when applied to classify

a large, regional-scale high-spatial resolution (HR, 1- 5 m) remotely sensed dataset. A geographic object based image analysis (GEOBIA) approach is used, because GEOBIA has been found to be particularly effective for classifying HR remotely sensed data (Blaschke, 2010; Ma et al., 2017). The remotely sensed data used includes 4-band color infrared 1 m National Agriculture Imagery Program (NAIP) orthoimagery, as well as 1 m light detection and ranging (LIDAR)-derived normalized digital surface model (nDSM) and intensity rasters.

2. Study Area and Data

2.1 Description of Study Area

The study site is located in the state of West Virginia, USA, between latitudes 79° 55' W and 79° 30' W and longitudes 39° 42' N and 39° 0' N, encompassing a multi-county area including Preston county, and portions of Monongalia, Taylor, Barbour, and Tucker counties (Figure 1). The total size of the study area is 260,975 ha, which is 4.2% of the area of the entire state of West Virginia. The terrain is mountainous, with elevations of 548 – 914 m, and mostly forested.

2.2 Remotely Sensed Data

Two types of remotely sensed data were utilized: passive optical multi-spectral imagery, and a LIDAR point cloud (Figure 2). The optical dataset comprises four-band color infrared leaf-on National Agriculture Imagery Program (NAIP) orthoimagery (Maxwell et al., 2017). The spectral bands of the NAIP imagery include red (590–675 nm), green (500–650 nm), blue (400–580 nm), and near-infrared (NIR) (675–850 nm) (Maxwell et al., 2014). The imagery has 1 m spatial resolution and 8-bit radiometric resolution. The NAIP data were acquired via a series of aerial flights between July 17 and July 30, 2011. A small portion of the imagery, less than 3% of the total NAIP dataset, was collected on October 10,

2011. The NAIP imagery were provided as 108 individual uncompressed digital orthophoto quarter quadrangles (DOQQs) in a tiff format.

The LIDAR data were acquired between March 28 and April 28, 2011, using an Optech ALTM-3100C sensor (WVU NRAC, 2013) with a 36° field of view and a pulse frequency of 70,000 Hz. The LIDAR

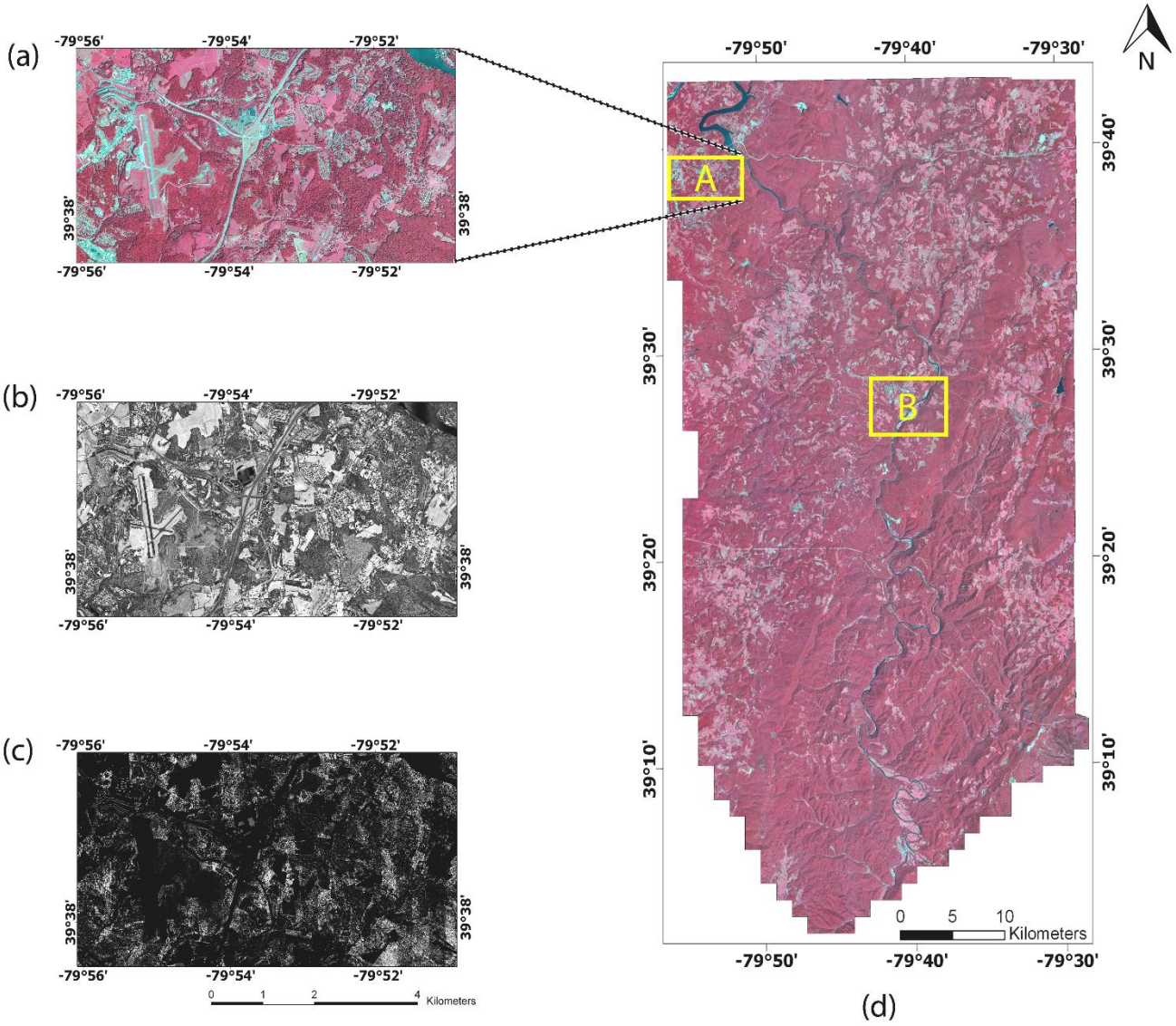

Figure 2 – Remotely sensed datasets: (a) small subset area, showing 4-band color infrared NAIP Orthoimagery, displaying bands NIR, Red, Green as RGB, (b) LIDAR-derived intensity, (c) LIDAR-derived nDSM, (d) false color composite NAIP orthomosaic displaying bands NIR, Red, Green, as RGB, of the entire study area. Highlighted areas "A" refers to the subset area represented in this figure, area "B" refers to a subset area chosen to display a sample classification area in Figure 5.

data were provided as 1164 individual .las files, containing a combined total of 5.6 x 10^9 points. The

LIDAR point cloud data include elevation, intensity, up to four returns, and a basic classification of the points provided by the vendor. A pilot investigation determined there was minimal change in the land cover during the approximately four months between the acquisition dates of most of the LIDAR and NAIP data.

2.3 Description of Land-Cover Classes

Four land-cover classes were mapped for this analysis, forest, grassland, water and other (Table 1).

Table 1 - Land-Cover Classes

Name	Description
Forest	Woody vegetation
Grassland	Herbaceous and other non-woody vegetation
Water	Man-made and natural waterbodies
Other	Bare soil, exposed rock, impervious surfaces, and bare croplands

3. Methods

3.1 Data Processing

The LIDAR tiles first were combined into a single large LIDAR point cloud file. Elevation and intensity information in the LIDAR point cloud were used to develop a normalized-digital surface model (nDSM) and an intensity raster, respectively. LIDAR-derived elevation and intensity surfaces have been demonstrated to improve the accuracy of land-cover classifications of HR multispectral imagery, especially if the spectral resolution of the imagery is low (Yan et al., 2015).

The LAS to Raster function in ArcMap 10.5.1 (ESRI, 2017) was used to rasterize the LIDAR point cloud. Elevation data in the LIDAR point cloud was used to first develop a bare earth digital elevation model (DEM) and a digital surface model from the ground and first returns, respectively. An nDSM was produced by subtracting the DEM from the DSM.

The first intensity returns in the LIDAR point cloud were rasterized to generate an intensity surface using the ArcMap LAS to Raster function. Slant range distance was not available, and thus it was not possible to correct for beam spreading loss or other factors. Previous research has shown that even in the absence of calibration of LIDAR intensity, LIDAR intensity data are useful for land cover classification (Maxwell et al., 2015). The pixel size of both the nDSM and LIDAR-intensity rasters was set to 1 m, matching the pixel size of the NAIP orthoimagery.

The NAIP tiles were mosaicked and color-balanced into a single large image using the Mosaic Pro tool in ERDAS Imagine 2014. As NAIP imagery comprise multiple flight lines of data acquired at different times of the day (Maxwell et al., 2017), radiometric variation can occur between NAIP tiles. In regional-scale analyses of NAIP data, this can be a particular concern, as a larger study area is likely to include more radiometric variation. Thus, color-balancing was applied during the mosaic process to reduce radiometric variation between the NAIP orthoimagery (Lear, 2005). The NAIP orthomosaic was then clipped to the boundaries of the LIDAR rasters. Layer stacking was then used to combine the NAIP and LIDAR rasters into a single layer stack containing six bands, four NAIP and two LIDAR.

3.2 Image Segmentation

Trimble eCognition Developer 9.3 multi-resolution segmentation (MRS) was chosen as the segmentation method for this analysis. MRS is a bottom-up region-growing segmentation approach that partitions images into distinct image segments (Baatz and Schäpe, 2002). Equal weighting was given to all six bands for the segmentation. Prior to the segmentation, a 5 x 5 pixel median filter was applied to the LIDAR data, as preliminary segmentation experiments indicated that the raw LIDAR data caused a large number of artefacts created in the image segmentation process. These artefacts were likely caused by the “sawtooth” scanning pattern of the OPTECH ALTM 3100 sensor and the 1 m rasterization process (Petrie and Toth, 2008).

The MRS algorithm has three parameters that require input from the analyst: scale, shape, and compactness (Baatz and Schäpe, 2002). The scale parameter (SP) determines the size of the image objects, and is usually assumed to be the most important (Belgiu and Drăguţ, 2014; Drăguţ et al., 2014; Kim et al., 2011). The SP value is typically chosen through trial and error (Arvor et al., 2013; Hay et al., 2005), although that approach has been criticized as *ad hoc,* not replicable, and not able to guarantee a near-optimal value (Kim et al., 2009). The estimation of scale parameter (ESP2) tool, an automated method for SP selection developed by Drăguţ et al. (2010), iteratively generates image-objects at multiple scale levels. The tool then plots the rate of change of the local variance (ROC-LV) against the associated scale parameter. Peaks in the ROC-LV curve indicate SPs with segment boundaries that tend to approximate natural and man-made features (Drăguţ et al. 2014).

As the ESP2 tool requires a large amount of computing resources, three small areas of the study area were randomly selected to run the ESP2 process. The results suggested SP values of 97, 97, and 104, and an intermediate value of 100 was selected for the MRS segmentation. The default shape and compactness parameters of 0.1 and 0.5 respectively were used, as varying these parameters did not appear to improve the quality of the segmentation. The segmentation of the regional-scale dataset generated 474,614 image segments.

3.3 Image Object Predictor Variables

Unlike pixels, which are uniform in size and shape, image segments in object-based image analyses can include spatial as well as spectral information. A total of 35 spectral and geometric predictor variables were generated for each image object (Table 2). Spectral variables include the mean, mode, standard deviation, and skewness of each image band. In addition, a separate spectral value *Brightness* was also included. The *Brightness* values of objects were calculated as the mean value of the four NAIP bands, averaged over all the pixels in the object (Salehi et al., 2012). NDVI was calculated

using the Red and NIR spectral bands from the NAIP data. Examples of geometric variables include object roundness, border length, and compactness.

Table 2 - Spectral and Geometric Attributes

Variable type	Object predictor variables	Number of variables
Spectral properties	Mean (Blue, Green, Intensity, NIR, Red, nDSM), Mode (Blue, Green, Intensity, NIR, Red, nDSM), Mean Brightness	13
Spectral Indices	Mean NDVI	1
Texture measures	Standard deviation (Blue, Green, Intensity, NIR, Red, nDSM), Skewness (Blue, Green, Intensity, NIR, Red, nDSM),	12
Geometric measures	Density, Roundness, Border length, Shape index, Area, Compactness, Volume, Rectangular fit, Asymmetry	9
Total		**35**

3.4 Sample Data Collection

Two large sample sets each containing 10,000 samples were collected from across the entire regional-scale dataset. One large sample set was used as training data while the other large sample set was used as an independent validation set used for testing the classifier accuracies, and was not used in training. As this analysis was conducted in a GEOBIA framework, image-objects were the sampling unit. Image objects were found to almost always represent a single class. In the rare instances that they did not, the majority class within the object was used as the class label.

Simple random sampling was used to acquire the validation dataset (Ramezan et al., 2019). Simple random sampling has the benefit that the population error matrix can be estimated directly from the sample statistics (Stehman and Foody, 2009). The size of the validation sample set (n = 10,000) was

approximately 2.1% of the regional-scale population. Image-objects were manually labeled by the analyst.

The training sample set were then collected from the regional-scale dataset. To ensure the independence of the validation dataset, the image-objects in the validation dataset were removed from the population before collecting the training dataset. As the study is overwhelmingly dominated by the forest class, the proportions of the classes in the image did not allow for an equalized stratified sampling. Therefore, disproportional stratified random sampling was used to ensure adequate representation of extreme minority classes in the training sets. Disproportional stratified random sampling involves the selection of samples from pre-defined strata, where each member of the stratum has an equal probability of being selected, but the size of the strata is defined by the analyst. Previous research has indicated that disproportional stratified random sampling is an effective approach for training data collection in regional-scale supervised land-cover classifications of HR remotely sensed data (Ramezan et al., 2019). Randomly collected training data improves the representativeness of the samples, and the disproportional stratified approach can reduce class imbalance (Ramezan et al., 2019). For this study, the strata sizes were defined as 50% Forest, 20% Grassland, 20% Other, and 10% Water.

The large training sample set (*n* = 10,000) (Figure 3) was randomly subset into a series of smaller training sets, with each subset independently chosen from the original 10,000, and each successive set approximately half the size of the preceding larger set. This resulted in training sets of size 10,000, 5,000, 2,500, 1,250, 626, 315, 159, 80, 40 (Class strata proportions were maintained for each sample set, which explains why each successively smaller sample is not exactly half of the larger set). The smallest training sample used was 40, because a preliminary analysis showed that sample sizes smaller than 40 caused problems with the cross-validation parameter tuning due to the small number of samples in the Water class. Table 3 summarizes the training sample sets and the validation dataset.

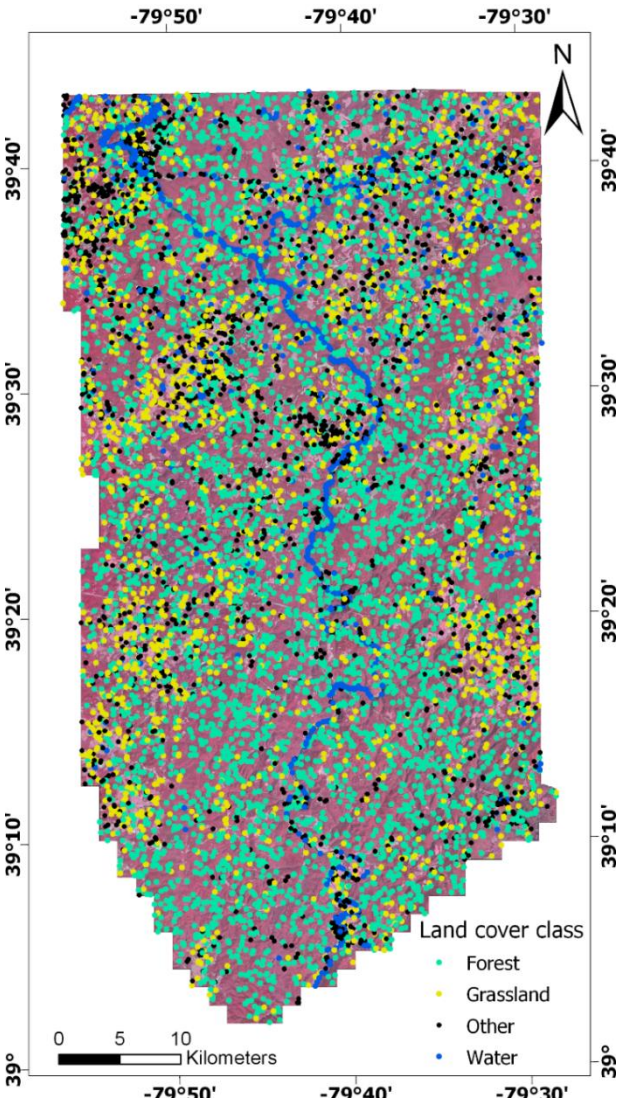

Figure 3 - Location of training sample polygons (n = 10,000). (Note: Each sample is indicated by a uniformed-size dot; the size and shape of the associated polygon is unrelated to the size and shape of the dot).

Table 3 - Training and validation data sample sizes

Purpose	Number of image-objects by class				Total sample size
	Forest	**Grass**	**Other**	**Water**	
Training	5,000	2,000	2,000	1,000	10,000
	2,500	1,000	1,000	500	5,000
	1,250	500	500	250	2,500
	625	250	250	125	1,250
	313	125	125	63	626
	157	63	63	32	315
	79	32	32	16	159
	40	16	16	8	80
	20	8	8	4	40
Validation	8,085	1,256	590	69	10,000

3.5 Supervised Classifications

Five supervised machine learning classifiers were compared in this study. The classifications were performed on each training dataset and evaluated against the regional-scale validation dataset. The classifications were performed within R 3.5.1 and R Studio 1.1.383; Table 4 lists the associated R packages.

Table 4 - List of R packages used

Machine Learning Classifier	Description	R Package	Reference
SVM	Radial basis function (RBF) kernel support vector machines	e1071 & caret	Meyer et al., 2012; Kuhn, 2016
RF	Fast implementation random forests suited for high dimensional data	ranger & caret	Wright and Zeigler, 2015; Kuhn, 2016
k-NN	Instance-based learning model using Euclidean distance	caret	Kuhn, 2016
NEU	Single-layer perceptron feed-forward neural networks	nnet & caret	Ripley and Venables, 2015; Kuhn, 2016
LVQ	Moving codebook vectors	class & caret	Ripley and Venables, 2015; Kuhn, 2016

3.5.1 Support Vector Machines (SVM) Classification

SVM is a non-parametric, supervised machine learning algorithm that seeks a hyperplane boundary to separate classes (Cortes and Vapnik, 1995; Pal, 2012). A distinctive feature of SVM is that the location of the hyperplane is determined by the training samples closest to the hyperplane, termed support vectors; other training samples are ignored. The optimization maximizes the margin of the hyperplane between the support vectors of the different classes, which is why SVM is sometimes referred to as a maximum margin classifier (Mountrakis et al., 2010). SVM is a binary classification method, so the classifier must be applied repeatedly for all possible combinations of classes. In addition, as the hyperplane is a linear decision boundary, and many classes are not linearly separable, SVM transforms the feature space to a higher dimension where the data may be linearly separable (Maxwell et al., 2018). This transformation is called the kernel trick. There is a variety of kernel types; we use a radial basis function kernel (RBF), a kernel commonly used in remote sensing (Maxwell et al., 2018; Mountrakis et al., 2010; Pal, 2012) and typically employed as a baseline for evaluating the performance of new SVM kernels (Sharma et al., 2016; Zhu and Blumberg, 2002).

3.5.2 Random Forest (RF) Classification

RF is an ensemble machine learning classifier that uses a large number of decision trees, each of which is given random subsets of the training data and predictor variables (Belgiu and Drăguţ, 2016). The decisions trees in the ensemble are produced independently and, unlike typical classification using a single decision tree, are not pruned. An internal cross-validation holdout process is used to estimate the performance of each tree in the forest. After training, each unknown sample is classified based on the majority vote of the ensemble (Kulkarni and Lowe, 2016). RF is a commonly used classification method in remote sensing analyses (Chen and Cheng, 2016; Gislason et al., 2004; Ramo and Chuvieco, 2017; Pal, 2005; Maxwell et al., 2018), and has become increasingly popular due to its superior classification accuracies compared to other commonly used classifiers such as single decision trees (Pal, 2005). Additionally, the RF classifier can be attractive to remote sensing scientists due to its ability to handle high dimensional datasets, an important consideration for hyperspectral and object-oriented datasets (Ham et al., 2005; Maxwell et al., 2018).

3.5.3 k-Nearest Neighbors (k-NN) Classification

The *k*-NN is a non-parametric classifier, which assigns class membership to new data inputs based upon their proximity to the *k* closest pre-labeled training data in the feature space. Lower *k*-values produce more complex decision boundaries, while larger *k*-values increase generalization (Maxwell et al., 2018; Everitt et al., 2011). *K*-NN is often described as a lazy learning classifier because it is not trained; unknowns are compared directly to the training data (Seetha, et al., 2012).

3.5.4 Neural Networks (NEU) Classification

NEU classifiers use a series of neurons, organized into layers. All neurons in neighboring layers are connected to each other by matrices of weights. Input layer neurons correspond to predictor variables, while output layer neurons correspond to classes (Maxwell et al., 2018). The neural network is trained by iteratively adjusting the weights to improve the classification, as the training data pass through layers. A feed-forward neural network with a single hidden layer is used in this analysis. Data in

this neural network moves only mono-directionally (forward), and uses only one single layer between the input and output layers (Ripley and Venables, 2016).

3.5.5 Learning Vector Quantization (LVQ) Classification

LVQ is a classifier that assigns membership to unseen examples using a series of codebook or prototype vectors within the feature space. Codebook vectors are typically randomly selected training data. Training samples in the LVQ algorithm are processed one at a time and are evaluated against the most similar codebook vector in the feature space. If the selected training sample has the same class as the codebook vector, a winner-take-all strategy is pursued, where the "winning" codebook vector is moved closer to the training sample. If the codebook vector does not have the same output as the training sample, the codebook vector is moved further away from the selected training sample in the feature space. This process is repeated until all codebook vectors have been evaluated against all training samples. Once the codebook vectors have been trained, the rest of the training data are discarded. The LVQ classifier predicts unseen examples in a similar manner to *k*-NN, except the codebook vectors are used for making predictions, rather than the full training data set. While LVQ is not commonly used in remote sensing analyses, it is a widely used classifier in many other fields because of its clear and intuitive learning process and ease of implementation (Grbovic and Vucetic, 2009).

3.6 Cross-Validation Parameter Tuning

Many supervised machine learning algorithms are parameterized, so they can be optimized for a specific objective or dataset (Karatzoglou et al., 2006). The selection of classifier parameters is an important stage of the classification process. However, as it is normally not possible to predict optimal values for these parameters, empirical cross-validation methods are typically employed (Brownlee, 2014; Heydari and Mountrakis, 2018; Ramezan et al. 2019).

Table 5 - Parameter tuning results

Classifier	Parameter	Parameter value by sample size								
	Sample Size	**40**	**80**	**159**	**315**	**626**	**1250**	**2500**	**5000**	**10000**
SVM	**Sigma**	0.0302	0.0343	0.0397	0.0342	0.0338	0.0305	0.036	0.0355	0.0355
	C	2	4	4	8	2	4	8	4	8
RF	**num.trees**	100	100	100	100	100	100	100	100	100
	mtry	9	20	13	13	13	13	28	24	13
k-NN	***k***	5	5	7	5	5	9	9	7	9
NEU	**size**	5	15	9	19	15	17	17	17	7
	decay	0.0075	0.1	0.0075	0.0001	0.0075	0.0178	0.1	0.1	0.1
LVQ	**size**	37	56	45	37	45	68	64	68	68
	k	1	6	6	6	6	11	1	6	31

K-fold cross validation testing was used for parameter tuning (Ramezan et al., 2019). Kappa was used to evaluate model parameters instead of overall accuracy, as several cross-validation models reported identical overall accuracy values, but different kappa coefficient values. Table 5 shows the optimal parameters for each classification.

After the optimal parameters for each classification were estimated, classifications were conducted for all five machine learning classifiers (SVM, RF, *k*-NN, NEU, LVQ), trained from each of the nine different sets, which varied in sample size (40, 80, 159, 315, 626, 1,250, 2,500, 5,000, 10,000 training samples), producing 45 separate classifications. Classifications were run on a custom workstation with an Intel Core i5-6600K Quad-Core Skylake processor and 32.0 GB of GDDR5 memory, and a Samsung 970 EVO NVMe 256 GB M.2. SSD running Windows 10 Pro. Processing time for all

classifications were recorded using the microbenchmark package (Ulrich, 2018). The processing time statistics are of interest as relative, and not absolute, values as processing time is highly dependent on a variety of factors such as system architecture, CPU allocation, memory availability, background system processes, among other factors.

3.7 Error Assessment

The classifications were evaluated against a large randomly sampled validation dataset consisting of 10,000 image-objects. As mentioned above, the pixels within image objects generally belonged to a single class. In the rare instances that more than one class was present in a single image object, the object was labeled based on the majority class. Results for each classification were reported in a confusion matrix. Overall map accuracy as well as user's and producer's accuracies were calculated, as well as the kappa coefficient. McNemar's test (McNemar, 1947) was used to evaluate the statistical significance of differences observed between classifications trained from sample sets of varying sizes. McNemar's test is a non-parametric evaluation of the statistical significance of differences between two classifications evaluated using related data. A *p*-value less than 0.05 specifies a one-sided 95% confidence that the differences in accuracy between the two classifications are statistically significant. McNemar's test was conducted between classifications using the same machine learning method, resulting in 180 individual tests. The full results of the McNemar's tests can be found in Appendix A.

4. Results and Discussion

Figure 4 summarizes the overall dependence of the five classification methods on sample size. RF is notable for consistently achieving higher overall accuracy than the other four machine learning algorithms, for all sample sizes. Fassnacht et al. (2014) also found that RF outperformed other classifiers such as SVM and *k*-NN, especially when large training sample sizes were used. RF saw its highest overall accuracy when trained from the 10,000 sample set (99.8%), and its lowest accuracy when the training sample size was only 40 (95.6%). However, the difference between the overall accuracy of these

classifications was only 4.2%, which was the second lowest difference between the highest and lowest performing classifications of all five machine learning classifiers, after LVQ, though the latter generally had low accuracy.

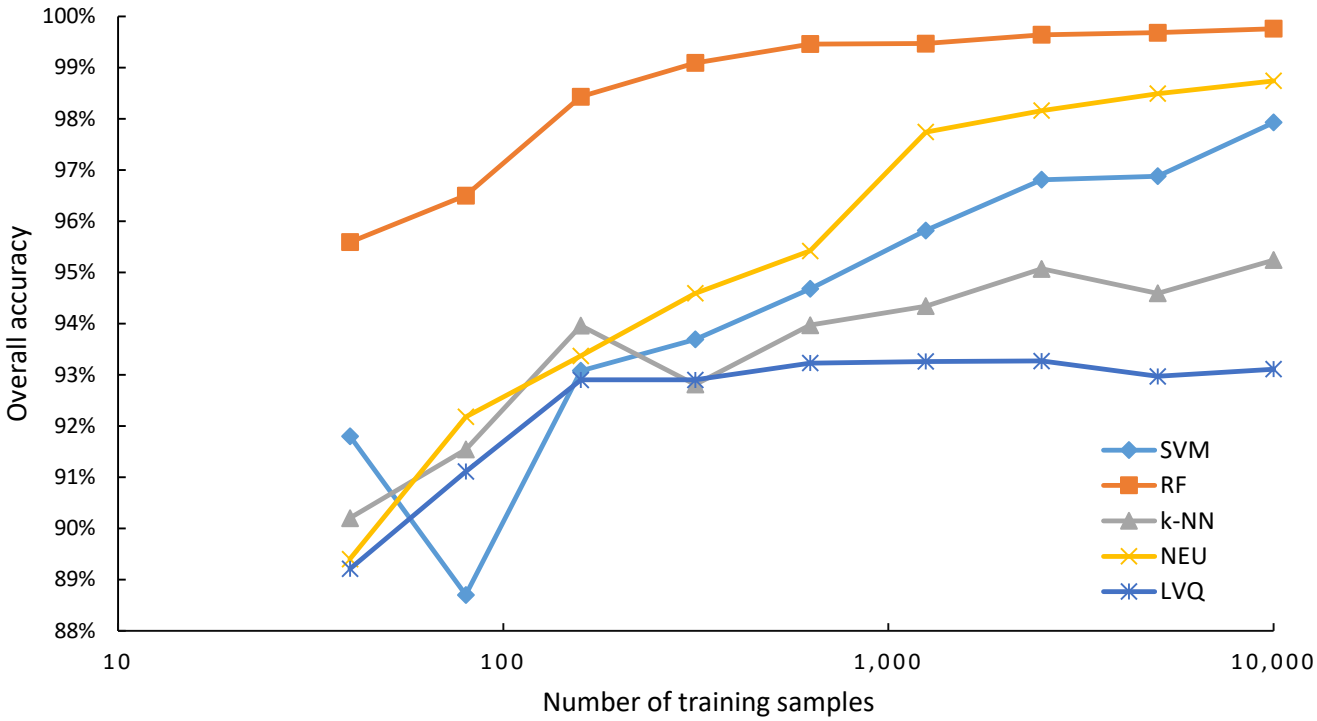

Figure 4 - Overall accuracy and training set size. Note that the x-axis is on a log-scale.

While the overall accuracy of the RF classifier increased as training sample size increased, the RF overall accuracy began to plateau when the sample size reached 626. The difference in accuracy between RF classifications using 626 and 10,000 samples was just 0.3%. Furthermore, the difference in accuracy for classifications between 2,500 and 5,000 samples was not statistically significant, though all other differences in accuracy for RF with different samples sizes were significant. This plateauing of the accuracy is perhaps not all that surprising; when classifications reach very high accuracy, there is little potential for further increases in accuracy.

The NEU classifier was generally the second most accurate classifier (Figure 4). NEU provided its highest accuracy (98.7%) when trained from the 10,000 sample set, and the overall classification accuracy decreased only 1.0% as the number of samples fell from 10,000 to 1,250 and 10,000. However, as the sample size was smaller than 1,250, the performance of the NEU classifier decreased rapidly. The lowest overall accuracy of all NEU classifications, 89.4% was the NEU classification trained from 40 samples. Of the five machine learning algorithms investigated in this study, NEU had the largest overall difference in accuracy between the classifications trained from 10,000 and 40 samples, 9.3%. It is notable that the threshold number of samples, below which the NEU classifier is most sensitive to sample size, was 1,250 samples or 0.26% of the population of the total study area. This threshold was very similar to the observation by Noi and Kappas (2017); and Colditz, (2015), who both found that overall accuracy of several machine learning classifiers began to decrease when sample size was smaller than a threshold of 0.25% of the study area. While Noi and Kappas (2017) observed this threshold was similar for three different machine learning classifiers, including RF, in this case, the thresholds for RF and LVQ classifiers are lower than that of NEU.

It should also be mentioned that while NEU performed very well when trained from large sample sets, almost equaling the performance of the RF, the NEU classifier was much slower than the other four classification methods. (Table 6) summarizes the runtime metrics for training and predicting processes for all classifications. The processing time for NEU is almost two orders of magnitude slower than RF, *k*-NN, and SVM, which is an important consideration in regional analyses, with very large datasets. Long processing times of NEU classifiers compared to other supervised machine learning algorithms was also noted in Maxwell et al. (2018).

Table 6 - Processing metrics (in seconds) for the five classifiers. Processing time ratio is the time for processing 10,000 samples as a multiple of the time for 40 samples.

Training Sample Size	*Processing time (s)*				
	SVM	RF	*k*-NN	NEU	LVQ
10,000	302.1	78.4	36.1	4379.4	2528.5
5,000	105.6	35.2	12.5	2231.4	738.1
2,500	33.7	16.7	4.5	1118.5	211.0
1,250	12.7	8.6	2.2	543.2	75.0
626	5.8	4.5	1.4	320.7	33.7
315	3.3	2.8	1.1	163.0	19.4
159	2.3	2.1	0.9	119.4	13.8
80	2.0	1.7	0.9	99.0	10.5
40	1.9	1.6	0.9	84.1	1.5
Processing time ratio	162	50	41	52	1744

The SVM classifier was typically the third best performing classifier when the sample size was between 315 and 10,000. When the sample size was increased from 5,000 to 10,000, SVM had the largest increase in performance, 1.1%, compared to the other four classifications methods, which saw increases of 0.1% to 0.7%. This is notable, as it suggests that SVM benefits from very large sample sets (n = 10,000), and does not plateau in accuracy as much as RF and NEU classifiers do when the sample size becomes very large (e.g. 10,000). This is likely due to larger samples containing more examples in the feature space that can be used as support vectors to optimize the hyperplane and thus identify a more optimal class decision boundary.

However, when trained from just 80 samples, SVM generated the lowest overall accuracy of all the classifications (88.7%), irrespective of the sample size or method. In contrast to the overall trend, reducing the sample size to 40 samples actually increased the accuracy of the SVM classifier to 91.8%. As indicated in Tables 7 and 8, the lower performance of SVM trained with 80 samples compared to SV trained with 40 samples was partly due the former classification's lower user's and producer's accuracies for grassland and lower producer's accuracy for forest. It is surprising that these two classes, the largest classes by area, should vary so in accuracy. However, since the samples are selected

randomly, and SVM focuses exclusively on support vectors (i.e. potentially confused samples) for separating classes, it suggests SVM may be inherently more inconsistent in its likely accuracy at any sample size. For the SVM trained with just 40 samples, the water class, which has only 4 training samples, resulted in the lowest producer's accuracy, just 40%. This is evident in Figure 5 (b), where SVM classification trained with 40 samples has many areas within the central river incorrectly mapped as other, unlike the SVM trained with 80 samples (Figure 5 (a)). These observations regarding SVM class accuracy further stress the value of larger sample sets for SVM classifications. Table 7 - Confusion Matrix for the SVM classification trained from 80 samples

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified Data (No. Objects)	**Forest**	7336	68	8	0	7412	99.0%
	Grassland	514	991	54	0	1559	63.6%
	Other	229	196	504	30	959	52.6%
	Water	6	1	24	39	70	55.7%
	Total	8085	1256	590	69	10000	**Overall Accuracy: 88.7%**
	Producer's Accuracy	90.7%	78.9%	85.4%	56.5%		

Table 8 - Confusion Matrix for the SVM classification trained from 40 samples

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified Data (No. Objects)	**Forest**	7655	153	42	1	7851	97.5%
	Grassland	277	1063	98	1	1439	73.9%
	Other	145	40	434	39	658	66.0%
	Water	8	0	16	28	52	53.8%
	Total	8085	1256	590	69	10000	**Overall Accuracy: 91.8%**
	Producer's Accuracy	94.7%	84.6%	73.6%	40.6%		

k-NN was the fourth best performing machine learning classifier for larger sample sizes, ranging from 626 to 10,000. While *k*-NN accuracy generally decreased as the sample size became smaller, *k*-NN was even more erratic in its overall trend than SVM. In two separate instances, *k*-NN classifications

trained from a smaller sample size produced a higher overall accuracy than *k*-NN classifications trained from double, or even triple, the sample size. For example, *k*-NN trained with 2500 samples had an overall accuracy of 95.1%, which was 0.6% higher than the overall accuracy of *k*-NN trained with 5000 samples. Similarly, the *k*-NN classification trained with 159 samples had a higher accuracy (94.0%) than when trained with 315 samples (92.8%), and an equivalent overall accuracy to when it was trained with 626 samples. This suggests that *k*-NN is somewhat inconsistent generalizer, and may be sensitive in random variations in the training data. One interesting observation with *k*-NN is that, despite being a lazy learning classifier, requiring each unknown to be compared to the original training data, it was consistently the fastest classifier, and furthermore, the processing time was the least affected by training sample size (Table 6). This is shown by the fact processing with 10,000 samples took only 41 times as long as with 40 samples. In comparison, RF took 50 times, SVM 162 times and LVQ 1,744 times as long.

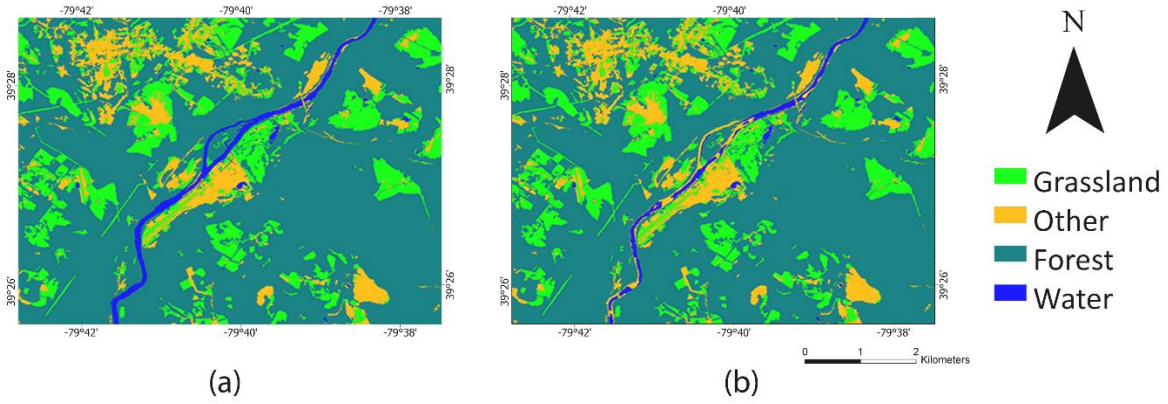

Figure 5 – Example of land-cover classifications: Sample classification area depicted in this figure is highlighted in Figure 1 (d) as area "B": (a) SVM trained with 80 samples, (b) SVM trained with 40 samples.

Except for SVM trained with 80 samples, LVQ had the lowest accuracy across all sample sizes. The performance gap between LVQ and the other four classifiers increased with sample size, because

the overall accuracy of LVQ plateaued when the sample size reached 159. Indeed, LVQ overall accuracy remained almost unchanged when the sample size increased from 159 to 10,000, with less than a 0.4% difference between all classifications, and none of the differences between classification accuracies trained on successively larger samples was statistically significant (Appendix A). Out of the five classification methods, LVQ also showed the lowest difference in overall accuracy between its most accurate classifications, those trained with 2500 and 1250 samples, which both had an accuracy of 93.3%, and the lowest performing classification, trained with 40 samples, at 89.2%. This indicates that LVQ is less sensitive to sample size than the other four classification algorithms. Nevertheless, LVQ processing time was the most sensitive to sample size, increasing 1,744 fold, as the sample size increased from 40 to 10,000 samples.

5. Conclusion

This analysis explored the effects of training sample size, varying from 40 to 10,000, on five supervised machine learning algorithms, SVM, RF, *k*-NN, NEU, LVQ, to classify a regional-scale HR remotely sensed dataset. Although it is well known from previous studies that larger training sets typically provide superior classification performance over smaller training sets (Foody et al., 2006), our study found considerable variation in how five machine learning classifiers responded to changes in training sample size. Furthermore, our study extends previous comparisons of machine learning classifier dependence on sample size (e.g., Qian et al., 2015 and Shang et al., 2018) by incorporating RF, SVM, NEU and *k*-NN, which tend to be amongst the most commonly used classifiers (Maxwell et al., 2018) as well as LVQ, a method widely used in non-remote sensing disciplines. We also evaluate performance over a very large range compared to previous studies.

Overall, given RF's superior performance over the other machine learning algorithms across all sample sizes, RF appears to be the best choice for constructing regional-scale land-cover classifications

using HR remotely sensed data. RF was found to be one of the most robust classifiers to decreases in training sample size, so much so, that it outperformed several other supervised classification approaches even when the other classifiers were given much larger training sets. In this case, RF using 315 samples provided a higher overall accuracy than the other four classification methods, even when given more than 30 times the number of training samples. Thus, RF is a particularly good choice if there are only limited training data. RF was also the second fastest classifier, after *k*-NN. Along with k-NN, and NEU, RF processing time was the least affected by increasing numbers of training samples (Table 6). Although this paper has not focused on the other attributes of RF, it is worth noting that RF offers the benefits of an estimate of classification accuracy (the so-called out-of-bag accuracy), and variable importance (Maxwell et al., 2018). Finally, RF is relatively straightforward to parameterize. In this study, we optimized only the number of decision trees (i.e, the num.tree variable), though the result is not particularly dependent on the number used, as long as it is large (Maxwell et al., 2018).

NEU was found to be generally the second-most accurate classifier, especially for classifications using a larger number of samples (315 and above). NEU was however was the classifier most affected by sample size. Therefore, if using NEU, it is particularly important to get the largest training sample data set possible. Unfortunately, though, NEU was also the slowest classifier, as much as 2 orders of magnitude slower than the other methods. Thus, even if a large training data set is available, making NEU accuracy competitive with that of RF, the slow processing time makes NEU unattractive.

SVM and *k*-NN were generally (though not always) the third and fourth-most accurate classifiers. Although they both showed the general pattern of declining accuracies with smaller sample sizes, the pattern was inconsistent, perhaps indicating that these two methods are more sensitive to the particular values of each training sample chosen, which is particularly a factor for smaller sample sizes. It is notable that the water class, which had the fewest samples, varied the most in accuracy in Tables 7

and 8. One benefit of *k*-NN is that it was the fastest classifier, and processing time was the least sensitive to the number of training samples.

LVQ was the method with overall accuracies least affected by sample size: overall accuracy essentially was unchanged for sample sizes of 159 and larger. Nevertheless, this method's processing time was the most affected by the number of training samples. Thus, for LVQ there is a strong disincentive to using a large sample.

In conclusion, machine learning methods vary considerably in their response to changes in sample size. Nevertheless, in this case, RF appeared to be the best all-round choice as a classifier. The insights provided by this analysis can be valuable for informing decision analysis on classifier selection given a particular training set size in future applied remote sensing projects.

References

Arvor, D., Durieux, L., Andrés, S., Laporte, M. 2013. Advances in Geographic Object-Based Image Analysis With Ontologies: a review of main contributions and limitations from a remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*. 82: 125-137. DOI: 10.1016/j.isprsjprs.2013.05.003.

Baatz, M. and Schäpe, A. 2000. Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. Strobl, T. Blaschke, G. Griesebner (Eds.), Angewandte Geographische Informations-Verarbeitung XII, Wichmann Verlag, Karlsruhe, Germany (2000) pp, 12-23

Belgiu, M. and Drăgut, L. 2014. Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*. 96: 67-75. https://doi.org/10.1016/j.isprsjprs.2014.07.002.

Belgiu, M. and Drăgut, L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*. 114: 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Blaschke, T. 2010. Object based image analysis for remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*. 65(1): 2-16. https://doi.org/10.1016/j.isprsjprs.2009.06.004.

Brownlee, J. 2016. Learning Vector Quantization for Machine Learning. https://machinelearningmastery.com/learning-vector-quantization-for-machine-learning/

(last date accessed: 18 Feb 2019)

Chen, L., and Cheng, X. 2016. Classification of High-Resolution Remotely Sensed Images Based on Random Forests. *Journal of Software Engineering*. 10(4): 318-327. DOI: 10.3923/jse.2016.318.327.

Chen, Gang., Hay, Geoffrey J., St-Onge, Benoit. 2011. A GEOBIA framework to estimate forest parameters from lidar transcts, Quickbird imagery and machine learning: A case study in Quebec, Canada. *International Jouranl of Applied Earth Observation and Geoinformation*, 15(2012): 28-37. http://dx.doi.org/10.1016/j.jag.2011.05.010.

Cortes, C., and Vapnik, V. 1995. Support-Vector Networks. Machine Learning 20: 273–297. doi:10.1007/BF00994018.

Colditz, R. R. 2015. An Evaluation of Different Training Sample Allocation Schemes for Discrete and Continuous Land Cover Classification Using Decision Tree-Based Algorithms. *Remote Sensing*. 7(8): 9655-9681. https://doi.org/10.3390/rs70809655.

Drăguţ, L., Tiede, D., Levick, S. R. 2010. ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*. 24(6): 859-871. DOI: 10.1080/13658810903174803.

Drăgut, L., Csillik, O., Eisank, C., and Tiede, D. 2014 Automated parameterization for multi-scale image segmentation on multiple layers. *ISPRS J. Photogrammetric Remote Sensing*. 88(100): 119-127. https://doi.org/10.1016/j.isprsjprs.2013.11.018

ESRI. 2017. ArcGIS Desktop: Release 10.5.1; Environmental Systems Research Institute: Redlands, CA, USA.

Everitt, B. S., Laundau, S., Leese, M., and Stahl, D. 2011. Miscellaneous Cluster Methods in Cluster Analysis, 5th Edition. John Wiley & Sons, Ltd, Chichester, UK.

Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernandez, J., Corvalan, P., Koch, B. 2014. Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sensing of Environment*. 154: 102-114. DOI:10.1016/j.rse.2014.07.028.

Foody, G. M., McCulloch, M.B., Yates, W. B. 1995. The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*. 16(9): 1707-1723. https://doi.org/10.1080/01431169508954507.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., Boyd, D. S. 2006. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment*. 1(15): 1-14. https://doi.org/10.1016/j.rse.2006.03.004.

Gislason, P.O., Benediktsson, J. A., and Dveinsson, J. R. 2004. Random Forest classification of multisource remote sensing and geographic data. *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*. Anchorage, AK, USA, 20-24, Sept. 2004. DOI: 10.1109/IGARSS.2004.1368591.

Grbovic, M., and Vucetic, S. 2009. Regression Learning Vector Quantization. (2009). *2009 Ninth IEEE International Conference on Data Mining*. Miami, FL, USA, December 28, 2009. DOI: 10.1109/ICDM.2009.145.

Gu, H. Y., Li, H. T., Yan, L., Lu, X. J. 2015. A framework for Geographic Object-Based Image Analysis (GEOBIA) Based on Geographic Ontology. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XL-7/W4, 2015*. 2015 International Workshop on Image and Data Fusion, Kona, Hawaii, 21-23.

Ham, J., Chen, Y., Crawford, M. M., Ghosh, J. 2005. Investigation of the random forest framework for classification of hyperspectral data. *IEEE Transactions Geosci. Remote Sensing.* 43(2005): 492-501.

Hay, G. J., Castilla, G., Wulder, M.A., Ruiz, J.R. 2005. An automated object-based approach for the multiscale image segmentation of forest scenes. *International Journal of Applied Earth Observation and Geoinfromation.* 7(4): 339-359. DOI: 10.1016/j.jag.2005.06.005.

Heydari, S. S., Mountrakis, G. 2017. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sensing of Environment.* 204. DOI: 10.1016/j.rse.2017.09.035.

Karatzoglou, A., Meyer, D., and Hornik, K. 2006. Support Vector Machines in R. *Journal of Statistical Software*. 15(9):1-28.

Kim, M., Warner, T.A., Madden, M., and Atkinson, D. 2011. Multi-scale texture segmentation and classification of salt marsh using digital aerial imagery with very high spatial resolution. *International Journal of Remote Sensing*. 32: 2825-2850.

Kim, M., M. Madden, and T.A. Warner 2009. Forest type mapping using object-specific texture measures from multispectral IKONOS imagery: Segmentation quality and image classification issues. *Photogrammetric Engineering and Remote Sensing* 75(7): 819-829.

Kuhn, Max. 2016. caret: Classification and Regression Training. R package version 6.0-71. https://CRAN.R-project.org/package=caret. (last date accessed: 18 Feb 2019).

Kulkarni, A. D., and Lowe, B. 2016. Random Forest Algorithm for Land Cover Classification. *International Journal on Recent and Innovation Trends in Computing and Communication.* 4(3): 58-63.

Lear, R.F. 2005. NAIP Quality Samples. United States Department of Agriculture Aerial Photography Field Office. Available Online: https://www.fsa.usda.gov/Internet/FSA_File/naip_quality_samples_pdf.pdf (last date accessed: 28 Dec 2018).

Li, Xiaoxiao and Shao, Guofan 2014. Object-Based Land-Cover Mapping with High Resolution Aerial Photography at a County Scale in Midwestern USA. *Remote Sensing*. 6: 11372-11390. doi:10.3390/rs61111372.

Ma, L., Li, M., Ma, X., Cheng, K., Du, P., Liu Y. 2017. A review of supervised object-based land-cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*. 130(2017):277-293. https://doi.org/10.1016/j.isprsjprs.2017.06.001.

Maxwell, A.E., Strager, M. P., Warner, T. A., Zegre, N.P., Yuill C. B. 2014. Comparison of NAIP orthophotragephy and RapidEye satellite imagery for mapping of mining and mine reclamation. *GIScience & Remote Sensing*. 51(3):301-320. https://doi.org/10.1080/15481603.2014.912874.

Maxwell, A.E., Warner, T.A., Strager, M.P., Conley, J.F., Sharp, A.L. 2015. Assessing machine-learning algorithms and image- and lidar-derived variables for GEOBIA classification of mining and mine reclamation. *International Journal of Remote Sensing*, 36(5): 954-978. http://dx.doi.org/10.1080/01431161.2014.1001086.

Maxwell, A. E., T.A. Warner, B.C. Vanderbilt, and Ramezan, C.A. 2017. Land cover classification and feature extraction from National Agriculture Imagery Program (NAIP) Orthoimagery: A review. *Photogrammetric Engineering and Remote Sensing* 83(11): 737-747. DOI: 10.14358/PERS.83.10.737.

Maxwell, A. E., Warner, T. A., Fang, F. 2018. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*. 39(9): 2784-2817. https://doi.org/10.1080/01431161.2018.1433343.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika.* 12(2): 153-157. DOI:10.1007/BF02295996.

Meyer, David. 2012. Support Vector Machines: The Interface to libsvm in package e1071. R package version 6.0-71. https://CRAN.R-project.org/package=e1071. (last date accessed: 18 Feb 2019).

Millard, K., and Richardson, M. 2015. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. *Remote Sensing* 7(7): 8489-8515. https://doi.org/10.3390/rs70708489.

Mountrakis, G., Im, J., Ogole, C. 2010. Support Vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*. 66(3): 247-259. https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Myburgh, G., Van Niekerk, A. 2013. Effect of feature dimensionality on object-based land cover classification: A comparison of three classifiers. *South African Journal of Geomatics*. 2: 13-27.

Noi, P. T., Kappas, M. 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*. 18(1): 18. https://doi.org/10.3390/s18010018.

Pal, M. 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing.* 26(1): 217-222. https://doi.org/10.1080/01431160412331269698.

Pal, M. 2012. Kernel Methods in Remote Sensing: A Review. *ISH Journal of Hydraulic Engineering*. 15(1): 194-215. https://doi.org/10.1080/09715010.2009.10514975.

Petrie, G., Toth, C.K. 2008. Airborne and Spaceborne Laser Profilers and Scanners. In Topographic Laser Ranging and Scanning: Principles and Processing; Shan, J., Toth, C.K., Eds.; CRC Press: Boca Raton, FL, USA, 2008.

Qian, Y., Zhou, W., Yan, J., Li, W., Han, L. 2015. Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery. *Remote Sensing*. 7(1): 153-168. https://doi.org/10.3390/rs70100153.

Raczko, E., and Zagajewski, B. 2017. Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *European Journal of Remote Sensing*. 50(1): 144-154. https://doi.org/10.1080/22797254.2017.1299557.

Ramezan, C. A., Warner, T. A., Maxwell A. E. 2019. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing*. 11(2): 185. https://doi.org/10.3390/rs11020185.

Ramo, R., and Chuvieco, E. 2017. Developing a Random Forest Algorithm for MODIS Global Burned Area Classification. *Remote sensing.* 9(11): 1193. https://doi.org/10.3390/rs9111193.

Ripley, B., and Venables, W. 2015. Package 'class': Various functions for classification, including k-nearest neighbor, Learning Vector Quantization and Self-Organizing Maps.

Ripley, B., and Venables, W. 2016. Feed-Forward Neural Networks and Multinomial Log-Linear Models. R package version 7.3-12.

Salehi, B., Zhang, Y., Zhong, M., Dey, V. 2012. Object-Based Classification of Urban Areas Using VHR Imagery and Height Points Ancillary Data. *Remote Sensing*. 4(8): 2256-2276. https://doi.org/10.3390/rs4082256.

Samaniego, L., and Schulz, K. 2009. Supervised Classification of Agricultural Land Cover Using a Modified *k*-NN Technique (MNN) and Landsat Remote Sensing Imagery. *Remote Sensing*. 2009(1): 875-895. DOI:10.3390/rs1040875

Seetha, M., Sunitha, K. V. N., and Devi, G. M. 2012. Performance Assessment of Neural Network and K-Nearest Neighbour Classification with Random Subwindows. *International Journal of Machine Learning and Computing*. 2(6): 844-847. DOI: 10.7763/IJMLC.2012.V2.250.

Shang, M., Wang, S., Zhou, Y., and Du, C. 2018. Effects of Training Samples and Classifiers on Classification of Landsat-8 Imagery. *Journal of the Indian Society of Remote Sensing*. 46(9): 1333-1340. https://doi.org/10.1007/s12524-018-0777-z.

Sharma, V., Baruah, D., Chutia, D., Raju, P., Bhattacharya, D. K. 2016. An assessment of support vector machine kernel parameters using remotely sensed satellite data. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT).* Bangalore, India. 20-21 May, 2016. DOI: 10.1109/RTEICT.2016.7808096.

Stehman, S.V.; Foody, G.M. Accuracy assessment. 2009. In: The SAGE Handbook of Remote Sensing; Warner, T.A., Nellis, M.D., Foody, G.M., Eds.; Sage Publications Ltd.: London, UK, pp. 129–145. ISBN 9781412936163.

Ulrich, J.M. 2018. Microbenchmark: Accurate Timing Functions. R Package Version 1.4-4. 2018. Available online: https://cran.r-project.org/web/packages/microbenchmark/microbenchmark.pdf (last date accessed: 18 Feb 2019).

Waldner, F., Jacques, D. C., and Low F. 2017. The impact of training class proportions on binary ropland classification. *Remote Sensing Letters*. 8(12): 1122-1131. https://doi.org/10.1080/2150704X.2017.1362124.

Wright, M. N., and Ziegler, A. 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. Journal of Statistical Software. 77: 1-17. DOI:10.18637/jss.v077.i01.

WVU NRAC. 2013. Aerial Lidar Acquistion Report: Preston County and North Branch (Potomac) LIDAR *.LAS 1.2 Data Comprehensive and Bare Earth. West Virginia Department of Environmental Protection. Available online: http://wvgis.wvu.edu/lidar/data/WVDEP_2011_Deliverable4/WVDEP_deliverable_4_Project_Report.pdf (accessed on 1 December 2018).

Yan, W. Y., Shaker, A., El-Ashmawy, N. 2015. Urban land cover classification using airborne LiDAR data: a review. *Remote Sensing of Environment.* 158: 295-310. DOI:10.1016/j.rse.2014.11.001.

Zhu, G., and Blumberg, D. G. 2002. Classification using ASTER data and SVM algorithms; the case study of Beer Sheva, Israel. *Remote Sensing of Environment*. 80(2): 233-240.

Appendix A

SVM									
SVM-10000	SVM-5000	SVM-2500	SVM-1250	SVM-625	SVM-315	SVM-159	SVM-80	SVM-40	
	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	SVM-10000
		0.6917	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	SVM-5000
			< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	SVM-2500
				< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	SVM-1250
					< 0.001	< 0.001	< 0.001	< 0.001	SVM-625
						0.02	< 0.001	< 0.001	SVM-315
							< 0.001	< 0.001	SVM-159
								< 0.001	SVM-80
									SVM-40

RF									
RF-10000	RF-5000	RF-2500	RF-1250	RF-625	RF-315	RF-159	RF-80	RF-40	
	< 0.001	0.007	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	RF-10000
		0.571	0.002	0.001	< 0.001	< 0.001	< 0.001	< 0.001	RF-5000
			0.01246	0.01038	< 0.001	< 0.001	< 0.001	< 0.001	RF-2500
				< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	RF-1250

					< 0.001	< 0.001	< 0.001	< 0.001	RF-625
						< 0.001	< 0.001	< 0.001	RF-315
							< 0.001	< 0.001	RF-159
								< 0.001	RF-80
									RF-40

k-NN									
k-NN-10000	k-NN-5000	k-NN-2500	k-NN-1250	k-NN-625	k-NN-315	k-NN-159	k-NN-80	k-NN-40	
	< 0.001	0.3612	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	k-NN-10000
		0.01127	0.214	0.003	< 0.001	0.007	< 0.001	< 0.001	k-NN-5000
			< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	k-NN-2500
				0.0644	< 0.001	0.09	< 0.001	< 0.001	k-NN-1250
					< 0.001	< 0.001	< 0.001	< 0.001	k-NN-625
						< 0.001	< 0.001	< 0.001	k-NN-315
							< 0.001	< 0.001	k-NN-159
								< 0.001	k-NN-80
									k-NN-40

NEU									
NEU-10000	NEU-5000	NEU-2500	NEU-1250	NEU-625	NEU-315	NEU-159	NEU-80	NEU-40	
	0.0388	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	NEU-10000
		0.02261	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	NEU-5000
			0.004	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	NEU-2500
				< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	NEU-1250
					< 0.001	< 0.001	< 0.001	< 0.001	NEU-625
						< 0.001	< 0.001	< 0.001	NEU-315
							< 0.001	< 0.001	NEU-159
								< 0.001	NEU-80
									NEU-40

LVQ									
LVQ-10000	LVQ-5000	LVQ-2500	LVQ-1250	LVQ-625	LVQ-315	LVQ-159	LVQ-80	LVQ-40	
	0.5656	0.517	0.5465	0.6353	0.3845	0.4568	< 0.001	< 0.001	LVQ-10000
		0.1856	0.2036	0.268	0.8118	0.8933	< 0.001	< 0.001	LVQ-5000
			0.9	0.8855	0.1446	0.1556	< 0.001	< 0.001	LVQ-2500
				0.9236	0.1551	0.1888	< 0.001	< 0.001	LVQ-1250
					0.182	0.2309	< 0.001	< 0.001	LVQ-625
						0.9317	< 0.001	< 0.001	LVQ-315
							< 0.001	< 0.001	LVQ-159
								< 0.001	LVQ-80
									LVQ-40

Chapter 4

Recursive Feature Elimination applied to Supervised Machine Learning Classification: Training samples size and the Hughes Phenomenon

Abstract

Many remotely sensed datasets, such as hyperspectral imagery or geographic object-based image analysis (GEOBIA) data, incorporate a large number of predictor variables. However, when training data are limited, reducing the data dimensionality generally improves the accuracy of the classification, a phenomenon commonly known as the Hughes phenomenon. Feature selection methods such as recursive feature elimination (RFE) are therefore sometimes used to reduce the data dimensionality by eliminating redundant, noisy, or other features that do not contribute to the performance of the classifier. While RFE has been used in a number of experiments, investigations examining RFE typically use a static sample size, or only examine a single supervised machine learning classifier. This analysis therefore investigates the effect of RFE on the accuracy of three commonly used supervised machine learning algorithms, support vector machines (SVM), random forests (RF), and single-layer perceptron neural networks (NEU), trained from sample sets of varying size, ranging from a very small sample (n = 40) to a very large sample (n = 10,000). A GEOBIA approach was adopted, using high spatial resolution passive optical multispectral National Agricultural Imagery Program (NAIP) orthoimagery and LIDAR-derived raster grids, covering a 2,609 km^2 regional-scale study area in northeastern West Virginia, USA. RFE was found to consistently improve the overall accuracy of SVM and NEU classifiers, with the highest performance increase at 5.1% for the SVM classification trained from 40 samples, however the benefit of feature selection diminished as sample size increased. Overall, RF consistently provided the highest overall accuracy for any training sample size, and even outperformed SVM and NEU trained with RFE-optimized feature sets. Furthermore, RFE resulted in only a small improvement of RF classification accuracy. Thus, RF was more robust to the Hughes phenomenon than SVM or NEU. In summary, feature selection can result in a notable improvement in classification accuracy, and should be included in best practices for machine learning remote sensing analyses.

1. Introduction

1.1 Supervised Machine Learning and the Hughes Phenomenon

Many remotely sensed datasets incorporate a large number of predictor variables or features (Huang et al., 2017). Hyperspectral imagery has, by definition, many bands. Geographic object based image analysis (GEOBIA) data also tend to have high feature dimensionality, for example, including spectral, spatial, geometric, and other properties. Although increasing the number of predictor variables has the potential to improve the separability of classes (Ramezan et al. 2019b), Hughes (1968) showed that as the number of dimensions in a feature space increases, the number of training samples needed to maintain statistical confidence for a parametric classifier also increases. Thus, for a fixed number of training samples, adding more and more variables will eventually cause the classification accuracy to decline (Alonso et al., 2011; Thenkabail et al., 2014). This phenomenon is often referred to as the *curse of dimensionality,* or the Hughes phenomenon, despite the fact that criticisms have been raised regarding Hughes' original formulation of the issue (Van Campenhout, 1978).

The Hughes phenomenon is important in remote sensing because training data are often limited, for example, if field observations are required, and the study area is large or transportation between sites is difficult. Reducing the number of predictor variables to minimize the likelihood of the Hughes phenomenon affecting the classification accuracy is a widely discussed topic in remote sensing literature (Warner et al., 1999). Two of the most popular remote sensing strategies for data dimensionality reduction are feature extraction and feature selection. Feature extraction involves a transformation of the original features to create a new feature set in which most of the useful information is represented by a smaller number of features. Feature selection, on the other hand, is the selection of a subset of the original bands that is assumed to carry most of the information. Principal component analysis (PCA) and linear discriminant analysis are commonly used feature extraction methods. For example, Imani and Ghassemian (2015) used feature extraction methods including linear

discriminant analysis to improve an SVM classification of hyperspectral data. While feature extraction methods have proven to be useful in some remote sensing analyses, feature selection is sometimes a preferred method over feature extraction as the transformed features created by feature extraction can be difficult to interpret (Yu, 2003).

1.2 Feature Selection Methods

Feature selection methods can be divided into three groups: filter, embedded, and wrapper methods. Filter methods generally use independent statistical tests such as chi-square or analysis of variance (ANOVA) tests to rank the importance of features based upon an independent, generalized performance criterion such as information gain, or Fisher score (Jović et al., 2015). Filter-based feature selection methods are computationally efficient compared to wrapper or embedded methods as they do not require iterative testing of machine learning classification models. However, as filter-based feature selection is conducted using metrics that are not necessarily relevant to the discriminant function of the subsequent classification, the derived feature sets are not necessarily optimal (Kaushik, 2016).

Embedded methods incorporate the feature selection process into the machine learning algorithm itself. Popular examples of embedded feature selection algorithms are LASSO and RIDGE regression (Kaushik, 2016), as well as classification and regression trees (CART) (Breiman et al., 1984). Although embedded methods have been used in remote sensing analyses, such as Yu et al.'s (2006) object-based classification of HR aerial imagery, Bittencourt and Clarke's (2003) land-cover classification of Landsat-TM and AVIRIS imagery, or Archibald and Fann's (2007) classification of hyperspectral AVIRIS imagery, embedded feature selection is conducted as part of the classification process, and is native to classification method. As embedded methods are classifiers in themselves, optimized feature sets derived using embedded methods cannot be used in other machine learning algorithms, and are therefore inherently limited in their general application. Thus, they are considered to be not as flexible as wrapper-based methods (Jović et al., 2015).

The third approach, wrapper-based methods, constructs predictive models via a machine learning classifier using subsets of the original feature set. Cross-validation techniques are used to hold out part of the training set to test the performance of the models. The model with the highest accuracy is assumed to be the optimum choice. Forward selection and backwards elimination methods such as recursive feature elimination (RFE) are examples of common wrapper-based feature selection methods. One drawback with wrapper-based methods is that because they involve iterative testing of different combinations of the feature set, they tend to be computationally intensive, especially if the dimensionality of the dataset is high. With constant advances in computing power (i.e. CPU clock speed, floating-point operations per second [flops]) this issue is becoming less of a concern.

Wrapper-based methods are popular because feature variables are evaluated using machine learning algorithms, and do not rely on a generic performance criterion, like filter-based methods. Furthermore, while wrapper-based methods use a machine learning classification for feature selection, unlike embedded methods, the optimized feature set is not tied to the classification method used in the feature selection process, and can be applied to a variety supervised machine learning algorithms (Jović et al., 2015). Thus, wrapper-based methods may be particularly attractive for remote sensing applications, given the wide variety of machine learning algorithms that can be used to classify remotely sensed datasets (Maxwell et al., 2018; Ramezan et al., 2019b).

This work therefore focuses on a wrapper-based feature selection method because of the wide potential application and high general reliability of the approach. Recursive feature elimination (RFE) (Kuhn, 2016) is used because it is adaptable to most machine learning methods. For example, Stevens et al. (2013) used random forest (RF) RFE to improve the prediction of soil organic carbon using reflectance data acquired from visible and near infrared spectroscopy. Guan et al. (2012) used RFE to improve a high resolution land-cover random forest classification using LIDAR data and orthoimagery. Colkesen and Kavzoglu (2016) used a SVM-RFE approach to increase the accuracy of a classification

using hyperspectral AVIRIS data. In addition to being a popular method for improving supervised classification accuracy, RFE is often used as a baseline feature selection method for evaluating new feature selection methods in remote sensing analyses. For example, Zhang and Ma (2009) used SVM-RFE to evaluate a new feature selection algorithm, modified recursive SVM (MR-SVM).

While RFE has been demonstrated to improve the accuracy of a number of supervised machine learning classifiers, and has been previously used for feature selection in HR remote sensing analyses, most studies that examine RFE and its effect on classification accuracy use a static sample size (Colkesen and Kavzoglu, 2016; Huang et al., 2014). One exception is Pal and Foody (2010), who examined the RFE process on a supervised classifier using multiple sample sizes. However, their range of sample size was relatively narrow, and only focused on a single supervised classifier, SVM.

1.3 Research Aims

This study investigates feature selection using recursive feature elimination, the Hughes phenomenon, and three machine learning classifiers, SVM, RF, and single layer perceptron neural networks (NEU). These issues are explored in the context of a regional-scale land-cover classification using training sets that range from n = 40 to 10,000 samples. While increasing sample set size can help to reduce the potential negative effects on classification accuracy due to the Hughes phenomenon, collecting additional training data may not be practical or possible on large, regional-scale applied remote sensing analyses. Thus, a thorough investigation of the relative benefits of feature selection on several common machine learning classifiers may help inform future applied analyses involving high dimensional datasets with limited training data and may provide a model for future studies.

This paper is part of a larger study on optimizing machine learning classification. In a previous study we investigated sampling strategies and cross-validation methods (Ramezan et al. 2019a), and demonstrated the benefit of a random sample over a deliberative sample for training data. The current paper extends the findings of Ramezan et al. (2019b), which focused on the sensitivity of classifiers to

sample size, by investigating the potential benefits of feature selection as a way to reduce classifier dependence on sample size.

2. Study Area and Data

2.1 Description of Study Area

This analysis focuses on a multi-county area in the state of West Virginia, USA, including Preston County, as well as portions of neighboring Monongalia, Taylor, Barbour, and Tucker counties. The study site is 260,975 ha in size (just over 4% of the state), and generally encompasses the area 79° 55’ W to 79° 30’ W and 39° 42’ N to 39° 0’ N (Figure 1). The area is forested, and the terrain mountainous.

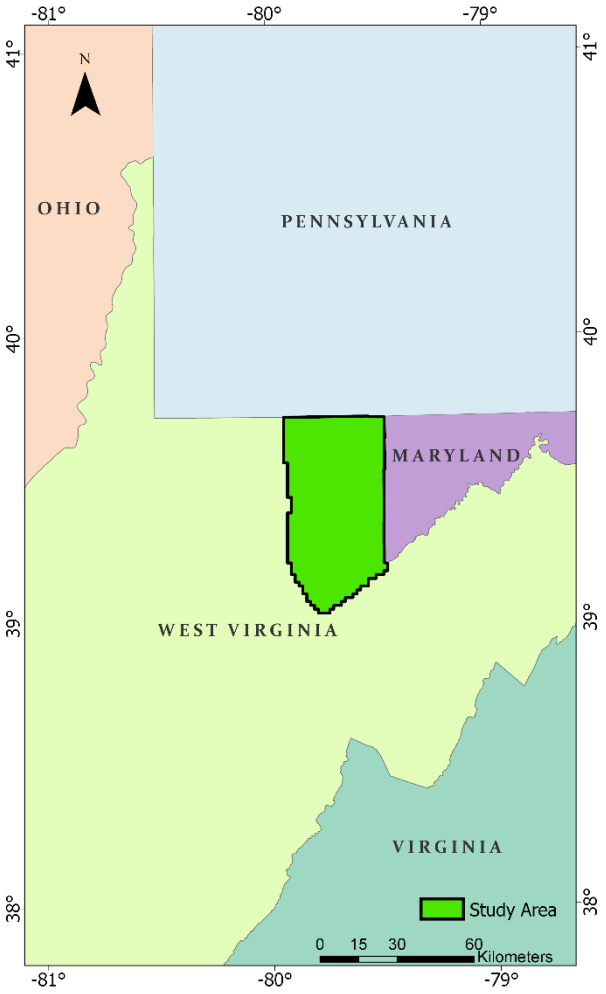

Figure 1 – Study Area

2.2 Remotely Sensed Data

Data from active and passive remote sensors were used. The passive optical dataset is four-band color infrared leaf-on National Agriculture Imagery Program (NAIP) orthoimagery (Figure 2) (Maxwell et al., 2017). The NAIP imagery comprises four spectral bands: red (590–675 nm), green (500–650 nm), blue (400–580 nm), and near-infrared (NIR) (675–850 nm) (Maxwell et al., 2014). The NAIP imagery has 8-bit radiometric resolution and 1 m spatial resolution.

The NAIP data were acquired between July 17 and July 30, 2011, although a small section of the imagery was collected on October 10, 2011. The NAIP data comprised 108 digital orthophoto quarter quadrangles (DOQQs), which were provided in an uncompressed .tiff format.

The active remote sensing dataset is aerial LIDAR point clouds acquired by an Optech ALTM-3100C sensor (WVU NRAC, 2013) between March 28 and April 28, 2011. The sensor has a 36° field of view and a laser pulse frequency of 70,000 kHz. The LIDAR data were provided as 1,164 individual .las point cloud files, which contained elevation, intensity, up to four returns, and a vendor-provided basic classification of the points. In total, the LIDAR data comprised 5.6 x10^9 points. A preliminary investigation determined that minimal change occurred during the approximately four months between the acquisition of the NAIP and LIDAR data.

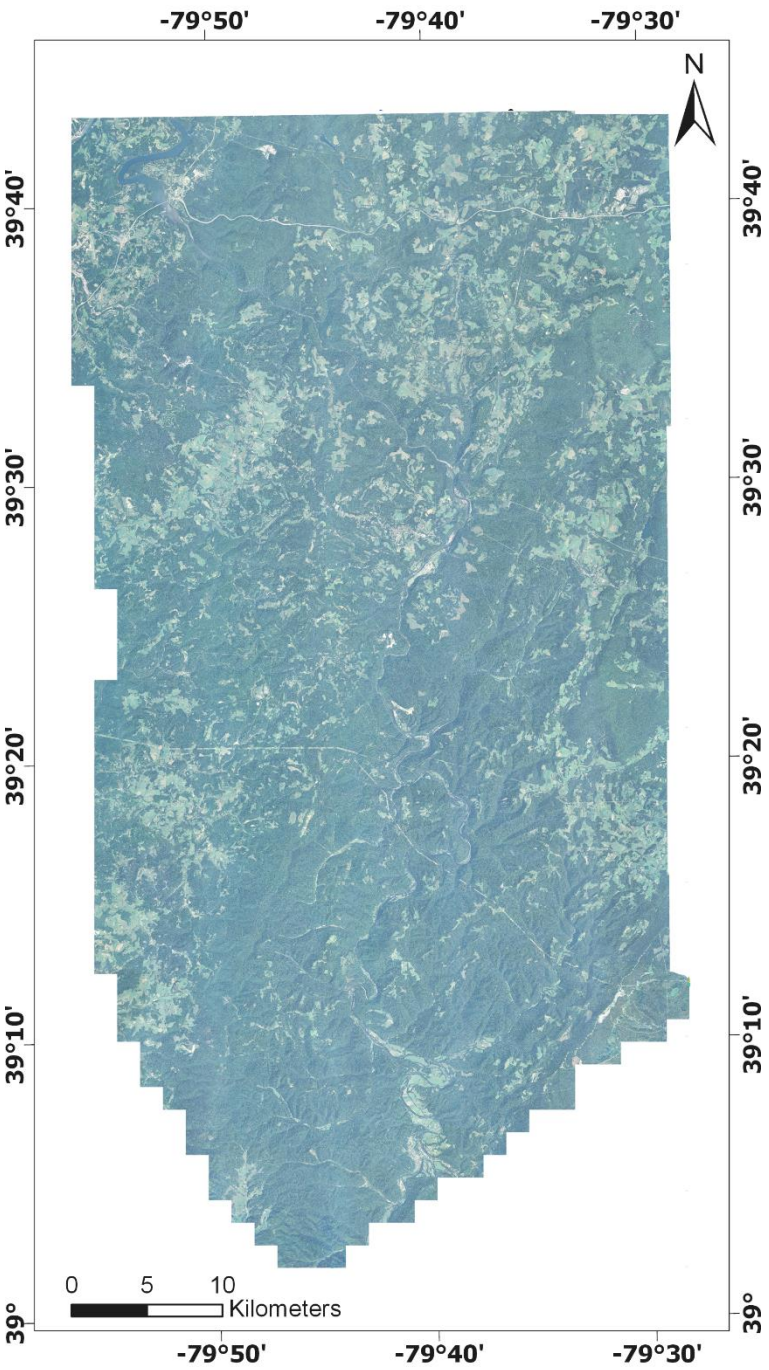

Figure 2 - NAIP orthomosaic of study area displaying bands Red, Green, Blue as RGB

2.3 Description of Land-Cover Classes

Four land-cover classes were mapped: Forest, Grassland, Water, and Other (Table 1).

Table 2 - Land Cover Classes

Class Name	**Description**
Forest	Appalachian mixed woody vegetation
Grassland	Non-woody vegetation
Water	Man-made and natural waterbodies
Other	Bare soil, exposed rock, impervious surfaces, and bare croplands

3. Methods

3.1 Experimental Design

Figure 3 provides an overview of the experimental design. Two separate sample sets, each comprising 10,000 samples, were collected. The first was set aside as a validation sample, and the second for training. The workflow includes feature selection, machine learning and accuracy assessment. Classifications with and without feature selection were undertaken with training sample sizes ranging from 40 to 10,000 samples. These steps are described in more detail below.

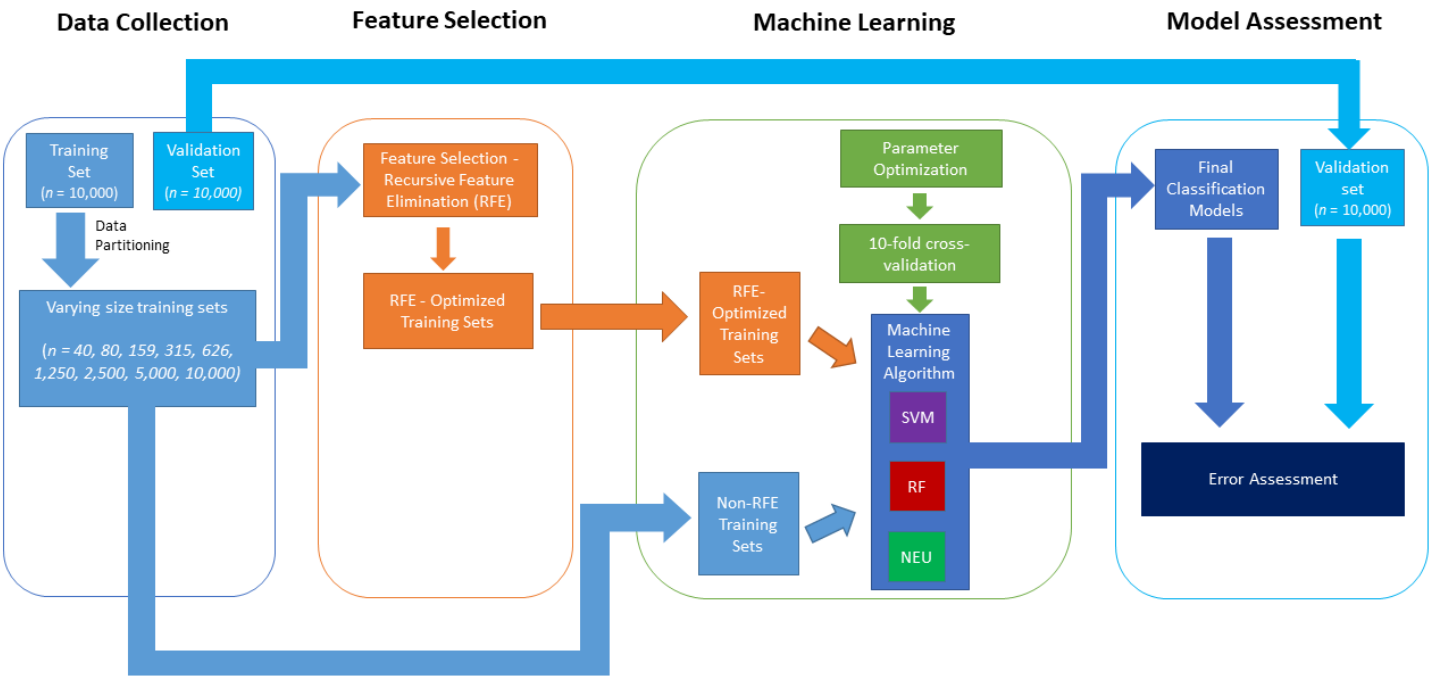

Figure 3 - Experiment Workflow

3.2 Data Processing

The 1,164 individual LIDAR point cloud tiles were first combined into a single LIDAR file. The LAS Dataset to Raster function in ArcMap 10.5.1 was used to generate rasterized datasets from the original point clouds. The rasterization was carried out at 1 m, which corresponds to the pixel size of the NAIP orthoimagery. Two intermediate products, a bare earth digital elevation model (DEM) and a digital surface model (DSM) were generated by rasterizing the ground and first elevation returns, respectively. A normalized digital surface model (nDSM), representing the heights of surface objects (mostly trees), was then calculated by subtracting the DEM from the DSM. The intensity raster dataset was developed by rasterizing the intensity of the first returns only. Only the nDSM and intensity raster datasets were used in the subsequent classification experiments.

All 108 NAIP orthoimages were color-balanced and mosaicked into a single large NAIP image using the Mosaic Pro tool in ERDAS Imagine 2016. Lear (2005) recommends color-balancing to reduce

radiometric variations between large NAIP data acquisitions acquired from different flights and times. After the mosaicking process, the NAIP mosaic was clipped to the boundaries of the LIDAR raster data. A layer stack process in ERDAS Imagine was then used to combined the NAIP and LIDAR raster data into a single image file containing six bands: Bands 1 – 4 comprised NAIP Red, Green, Blue, NIR and bands 5 & 6 the LIDAR nDSM and Intensity, respectively.

3.3 Image Segmentation

Segmentation was carried out with the multi-resolution segmentation (MRS) algorithm in Trimble eCognition Developer 9.3. MRS is a bottom-up region-growing segmentation. The MRS algorithm is one of the most commonly use segmentation methods in image processing and GEOBIA analyses (Kavzoglu and Tonbul, 2018). All six bands were weighted equally for the segmentation.

An initial test of the segmentation process found that a large number of artefacts, segments unrelated to any obvious ground features, were created. These artefacts appear to have been caused by a combination of the "sawtooth" scan pattern (Petrie and Toth, 2008) of the LIDAR sensor and the 1 m rasterization process. However, a 5 x 5 pixel median filter applied to both LIDAR-derived rasters grids before segmentation largely suppressed the problem.

Three user-defined parameters must be specified for the MRS algorithm: scale, shape, and compactness (Baatz and Schäpe, 2002). The scale parameter (SP) is usually found to be the most important of the three parameters, as it defines the average area of the image objects. (Belgiu and Drăguţ, 2014; Drăguţ et al., 2014; Kim et al., 2011). While *ad hoc* trial and error methods are typically used in most projects for selecting the SP (Arvor et al., 2013; Hay et al., 2005), such methods present challenges for experimental replication. We therefore used the estimation of scale parameter (ESP2) tool developed by Drăguţ et al. (2014), which can be considered an objective method for selecting the SP for the segmentation. The ESP2 tool iteratively generates image-objects at a range of scale levels.

The tool then plots the rate of change of the local variance (ROC-LV) against the associated scale parameter. Peaks in the ROC-LV curve are assumed to indicate SPs where image-object boundaries generally coincide with real-world objects, and thus represent optimal SP values for the segmentation.

Due to the demanding processing and memory requirements of the ESP2 tool, three randomly selected small subset areas were identified for estimating a representative SP value for the scene. The optimal scale parameters for the three subsets were 97, 97, and 104. An intermediate value of 100 was therefore chosen as the SP for the MRS segmentation of the entire dataset. In total, the MRS process produced 474,614 image segments.

3.4 Image-Object Attribute Feature Set

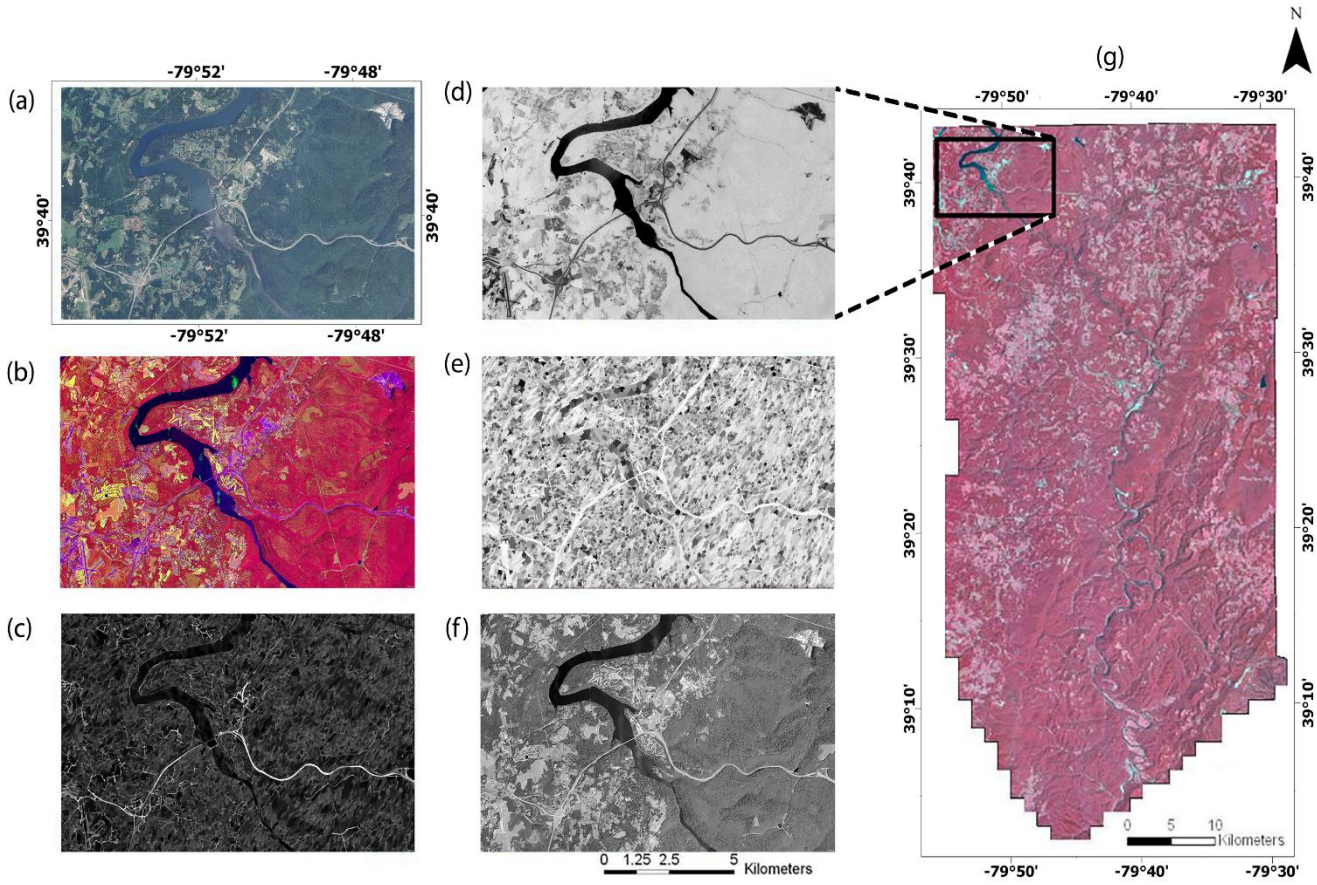

Figure 4 – Example images of the data, illustrating various image object attributes. (a) – (f) is an example subset. (g) is the entire study area. (a) and (g) are 4-band color infrared NAIP orthoimagery, (a) bands Red, Green, Blue as RGB, (g) bands NIR, Red, and Green as RGB, (b) false color composite image with NAIP-NIR, LIDAR-Intensity, NAIP-Red as RGB, (c) image-object compactness, (d) NDVI, (e) image-object asymmetry, (f) image-object Brightness.

A total of 35 attributes were generated for the image objects and used as predictor variables (or "feature set") for the classification (Table 2 and Figure 4). Spectral attributes for each object include the mode for each NAIP band, and textural variables, such as the standard deviation for each band. LIDAR-variables include the mean and mode of the LIDAR bands. Geometric attributes include object border length, and asymmetry. Additionally, the mean normalized difference vegetation index (NDVI) and mean Brightness (defined as the mean over all the NAIP bands) were calculated. With the exception of the mean NDVI, all spectral and geometric variables were generated using functions native to Trimble eCognition Developer 9.3.

Table 2 - Image-object attributes

Attribute Type	Attributes	Number of Attributes
Spectral variables	Mean (Blue, Green, NIR, Red), Mode (Blue, Green, NIR, Red)	8
Textural variables	Standard deviation (Blue, Green, Intensity, NIR, Red, nDSM), Skewness (Blue, Green, Intensity, NIR, Red, nDSM),	12
LIDAR variables	Mean (Intensity, nDSM), Mode (Intensity, nDSM)	4
Geometric variables	Density, Roundness, Border length, Shape index, Area, Compactness, Volume, Rectangular fit, Asymmetry	9
Spectral Indices	Mean NDVI Brightness	2

3.5 Training and Validation Sample Selection

The sampling unit for both the training and validation datasets were image-objects. Training and validation data were collected independently. Each data set comprised 10,000 samples, and represented approximately 2.1% of the total number of image-objects in the population. The validation dataset was collected using simple random sampling. Foody (2017) points out that simple random sampling is an effective design for validation datasets because the sample statistics directly estimate the population, and do not require additional manipulation prior to generating summary statistics.

The procedure for the training data were somewhat more complex than for the large, purely random validation data set, and drew upon previous research into sample size selection (Ramezan et al. 2019b). To ensure no overlap between the validation and training sets, all image-objects in the validation dataset were removed from the population before the training data were collected. As the dataset is dominated by the Forest class, and the other classes comprise a much smaller area, a purely random training sample would result in an imbalanced training set, which would likely reduce the classification accuracy (Ramezan et al., 2019b). Therefore, disproportional stratified random sampling, where the proportions in each strata are chosen by the analyst, was used. Disproportional stratified random sampling allows the analyst to boost the proportion of samples from classes that are too small to allow equally sized strata (Ramezan et al., 2019a). In order to apply this method, an initial classification is needed. A preliminary classification was therefore developed via a rule-based expert-system. The rule-based classification contained 16 individual rules. The accuracy of the rule-based classification was evaluated by the validation dataset, and had an overall accuracy of 98.1%. The class strata sizes for each class in the rule-based classification were found to be 79.8% Forest, 15.5% Grassland, 4.1% Other, and 0.6% Water. The disproportional strata sizes were defined as 50% Forest, 20% Grassland, 20% Other, and 10% Water. These strata sizes were selected to maintain the majority of

the Forest class, while ensuring a larger representation of other minority classes in the training set, especially the Water class, which was an extreme minority.

The training dataset was then randomly subset, while maintaining class proportion size, into a number of smaller training sets. Each training set was half the size of the preceding larger training set, resulting in training sets ranging in size from 10,000, 5,000, 2,500, 1,250, 626, 315, 159, 80, 40. The number of training samples in each sample set are summarized in Table 3.

Table 3 - Training and validation data sample sizes

Purpose	Sample selection method	Number of image-objects by class				Total sample size
		Forest	Grass	Other	Water	
Training	Disproportional random sample	5,000	2,000	2,000	1,000	10,000
		2,500	1,000	1,000	500	5,000
		1,250	500	500	250	2,500
		625	250	250	125	1,250
		313	125	125	63	626
		157	63	63	32	315
		79	32	32	16	159
		40	16	16	8	80
		20	8	8	4	40
Validation	Simple random sample	8,085	1,256	590	69	10,000

3.6 Feature Selection - Recursive Feature Elimination

The feature selection process used was Recursive Feature Elimination with resampling (RFE), which is built in to the caret package (Kuhn, 2016) in R. As described by Kuhn (2016), the RFE process is a wrapper-type method that uses backwards elimination and cross-validation resampling to rank

variables in order of importance to model accuracy and identify the best combination of features for a given sample set. The RFE process involves two distinct but connected stages. The first stage involves determining feature variable rankings by importance to classification accuracy, while the second stage involves determining the optimal number of feature variables to include in the feature set.

The algorithm begins by partitioning the sample set into training and test data via 10-fold cross-validation. All 35 feature variables were first used to train a 500-tree random forest classification as a baseline of accuracy for the classifier incorporating all feature variables. Backwards elimination is then used to iteratively remove different feature variables from the feature set, which are then used to train and test the effects of removed feature variables on the accuracy of the RF classification. The results are averaged across all folds. Overall accuracy was used as the performance metric for variable ranking. Thus, feature variables that contribute more towards the overall accuracy of the classifier are ranked higher than variables that do not contribute as much to the overall accuracy of the classifier. The result is a complete ranking of all 35 feature variables (Kuhn, 2016).

The variable rankings are preserved, and input into the second phase of the RFE process. The second phase of the RFE process first partitions the dataset into training and test data via 10-fold cross-validation. The full variable-ranked feature set is used to train a 500-tree RF classifier to determine a baseline level of accuracy. A second backwards elimination process is used to systematically remove variables from the ranked feature set. Variables are removed in each iteration from the bottom-up (the lowest ranked variables are removed first) in blocks of 5. 5 was chosen as the block size as it was the smallest number by which 35 the total number of feature variables, is divisible. The variable-size truncated feature sets are each used to train a 500-tree RF classifier for each fold. The results of each feature set classification were averaged across all folds. The feature set with the highest overall accuracy was chosen as the optimal feature set. Thus, the end result of both processes provides an optimally-sized ranked feature set.

The RFE process was conducted for each training set size, resulting in 9 RFE-derived feature sets. RFE-derived feature variable sets were stored separate from the original "No-RFE" feature sets, which contain the original 35 variables.

3.7 Parameter Optimization – *k*-fold Cross-Validation

Some supervised machine learning algorithms contain parameters which can influence the learning process of the classifier. While parameters can be manually specified, empirical methods such as cross-validation are often used to determine optimal parameters for a specific classifier and training set (Ramezan et al. 2019a). *K*-fold (*k* = 10) cross-validation was used to determine the optimal parameters for the three supervised classification methods when trained from each of the varying training sample sizes, as well as the RFE-optimized training sets (Table 4). The Kappa statistic was used as the accuracy metric to determine the optimal parameters, thus the parameters from the model with the highest Kappa value was determined to be the optimal parameter.

Table 4 - Parameter tuning results

Classifier	Sample Size	40	80	159	315	626	1,250	2,500	5,000	10,000
SVM	**Sigma**	0.0302	0.0343	0.0397	0.0342	0.0338	0.0305	0.036	0.0355	0.0355
	C	2	4	4	8	2	4	8	4	8
SVM-RFE	**Sigma**	0.1569	0.3381	0.0667	0.0835	0.0729	0.2813	0.036	0.06	0.0937
	C	4	8	4	2	1	4	8	8	16
RF	**num.trees**	100	100	100	100	100	100	100	100	100
	mtry	9	20	13	13	13	13	28	24	13
RF-RFE	**num.trees**	100	100	100	100	100	100	100	100	100
	mtry	2	2	7	7	4	3	36	18	15
NEU	**size**	5	15	9	19	15	17	17	17	7
	decay	0.0075	0.1	0.0075	0.0001	0.0075	0.0178	0.1	0.1	0.1
NEU-RFE	**size**	9	7	3	11	15	17	15	7	9
	decay	0.0031	0.0006	0.0075	0.0075	0.0422	0.0001	0.0178	0.1	0

3.8 Supervised Classification

The supervised machine learning classifiers in this study were SVM, RF, and NEU. The SVM classifier used a radial basis function kernel (RBF), a commonly used kernel in remote sensing (Meyer et al., 2012). An RF implementation particularly suited for high dimensional data was used (Wright and Zeigler, 2015). The NEU classifier was a feed-forward neural network containing a single hidden layer (Ripley and Venables, 2016).

After the optimal parameters for each classification were determined, classifications were conducted for all three machine learning classifiers (SVM, RF, and NEU) with one set of classifications "No-RFE" trained from the varying size sample sets that included the original full feature set of 35 variables, and one set of classifications "RFE" which were trained from the RFE-optimized feature sets. In total, 54 classifications were produced (9 training set sizes x 3 machine learning classifiers x 2 feature set types). The classifications were conducted within the R statistical client and associated packages: e1071 package for SVM (Meyer et al., 2012), ranger package for RF (Wright and Zeigler, 2015), and the nnet package for NEU (Ripley and Venables, 2016). The caret package (Kuhn, 2016) was used as the framework for processing the classifications and the accuracy assessments. Classifications were run on a custom workstation with an Intel Core i5-6600K Quad-Core Skylake processor and 32.0 GB of GDDR5 memory, and a Samsung 970 EVO NVMe 256 GB M.2. SSD running Windows 10 Pro.

3.9 Error Assessment

The classifications were evaluated against the randomly sampled validation 10,000 image-object dataset. Classification results were reported in a confusion matrix. Overall, user's, and producer's accuracy were calculated along with the Kappa coefficient. In addition to these accuracy measures, McNemar's test (McNemar, 1947) was used to evaluate the statistical significance of differences between No-RFE and Post-RFE classifications. McNemar's test is a non-parametric evaluation of the

statistical significance of differences between two classifications evaluated using related data. A *p*-value less than 0.05 specifies a one-sided 95% confidence that the differences in accuracy between the two classifications are statistically significant. McNemar's test was conducted between each No-RFE and Post-RFE classification of the same machine learning classifier trained from the same sample size, resulting in 27 individual tests.

4. Results and Discussion

RFE indicated that the optimal number of features was less than the original 35 features, for every training sample size except 2,500 (Table 5). The optimal number of features varied depending on the sample size; smaller numbers training samples generally (though not consistently) had a smaller optimal number of features. For example, for 5,000 and 10,000 training samples had 25 optimal features, whereas 80 training samples had only 5 optimal features.

The top five ranked variables for each sample set are listed in Table 6. Generally, the RFE process identified similar highly-ranked variables for each sample size for the training data, which gives some confidence in the reproducibility of the RFE approach, even as the number of samples varied. Notably, the two LIDAR variables, Mean intensity and Mode intensity, along with Mean NDVI all occur in all but one of the rankings. While geometric variables were mostly excluded by the RFE process in this analysis, geometric variables may provide value in other land-cover classifications (Guo et al., 2007). While the spectral index, Brightness, was retained for all sample sizes, except 1,250, 315, and 80, the RFE process did not rank the variable highly, indicating it is less important compared to other spectral variables or spectral indices such as Mean Intensity and Mean NDVI, respectively.

Table 5 - Optimal number of feature variables as determined by RFE

Training Set Size	Optimal Number of Feature Variables	Spectral	Textural	LIDAR	Geometric	Spectral Indices
10,000	25	8	9	4	2	2
5,000	25	8	9	4	2	2
2,500	35	8	12	4	9	2
1,250	10	4	2	3	0	1
626	20	7	7	3	1	2
315	10	4	2	3	0	1
159	25	8	9	4	2	2
80	5	1	1	2	0	1
40	15	4	5	4	0	2

Table 6 - Top five feature variables per sample size (LIDAR-derived variables are in shades of green, spectral indices are indicated in pink, NAIP-derived variables are in shades of orange)

	Sample Set Size								
Feature Variable Rank	**10000**	**5000**	**2500**	**1250**	**625**	**315**	**156**	**80**	**40**
1	Mean Intensity	Mean Intensity	Mean Intensity	Mean Intensity	Mean NDVI	Mean Intensity	Mean NDVI	Mean NDVI	Mean NDVI
2	Mean NDVI	Mean NDVI	Mean NDVI	Mean NDVI	Mean Intensity	Mean NDVI	Mean Intensity	Mean Intensity	Standard deviation nDSM
3	Mean NIR	Mean NIR	Mean NIR	Mode Intensity	Standard deviation nDSM	Mode Intensity	Standard deviation nDSM	Mean Red	Mean Intensity
4	Mode Intensity	Mode Intensity	Mode Intensity	Standard deviation nDSM	Mode Intensity	Standard deviation nDSM	Mode Intensity	Standard deviation nDSM	Mean nDSM
5	Standard deviation nDSM	Standard deviation nDSM	Standard deviation nDSM	Mean nDSM	Mean nDSM	Skewness Intensity	Mean nDSM	Mean nDSM	Mode Intensity

Figures 5, 6, and 7, summarize the No-RFE and RFE classification accuracy by sample size for each supervised machine learning algorithm. The RFE-optimized feature sets consistently provided superior or comparable accuracy for all three supervised machine learning algorithms. Notably, as sample size decreased for all three classifiers, the difference in performance between No-RFE and RFE supervised

classifications increased. This finding is consistent with the Hughes phenomenon, and indicates feature selection is most beneficial in improving overall accuracy when the number of training samples is small.

RFE improved the accuracy of the SVM classifier for all training sample sizes, and the results were statistically significant at the 0.05 level (Table 7). The RFE SVM classification trained from 40 samples (Figure 8) saw the largest improvement in overall accuracy, an improvement of 5.1%. After applying the RFE process, the SVM classifier required fewer training samples to achieve similar levels of overall accuracy compared to classification without RFE. Even when the training sample size was extremely large (n = 10,000), the RFE process improved the overall accuracy of the SVM classification by 0.5%, from 97.9% to 98.4% (representing a relative reduction of approximately 25% of the error). This indicates that SVM may be sensitive to the dimensionality of the dataset, and confirms that SVM can be affected by the Hughes phenomenon, even with large sample sizes. Previous research has been contradictory: Pal and Foody (2010) also found SVM susceptible to the Hughes phenomenon, whereas Melgani and Bruzzone (2004) did not observe a Hughes effect, even when large training sets were used.

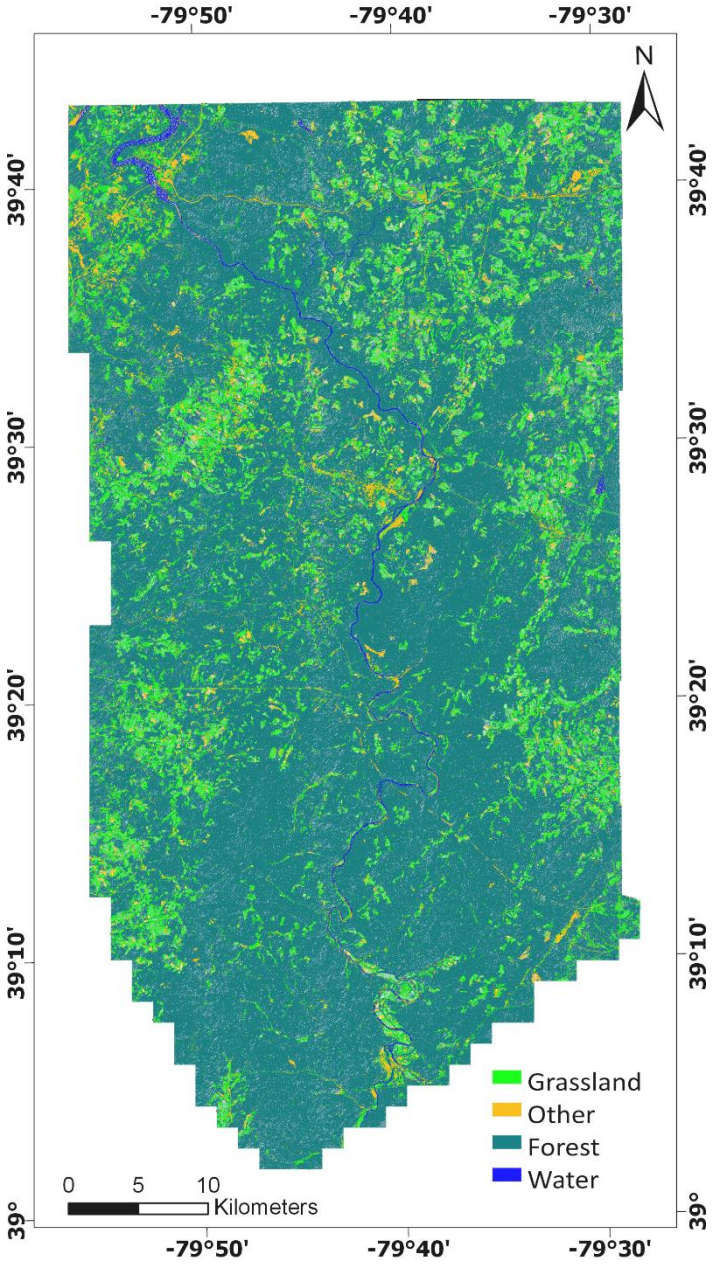

Figure 8 - Land-cover classification of SVM-RFE trained from 40 samples

RFE also improved the overall accuracy of the NEU classifier, but the results were more inconsistent than with SVM, and the benefits much smaller when the training set was large. When trained from 40 samples (Table 8), the RFE process improved the overall accuracy of the NEU classifier from 89.4% to 93.5%, an improvement of 4.1% (Table 9). Most notably, with RFE, the user's accuracy of Water and Grassland, two relatively rare classes, showed large increases, although this was partially offset by declines in producer's accuracy for these classes (Tables 8 and 9). RFE slightly improved overall accuracy of classifications with extremely large sample sets. For NEU with training sample sizes 5,000 and 10,000, the RFE process improved the overall accuracy of the NEU classifier by 0.4%. This is notable as it suggests that feature selection methods such as RFE may be beneficial to NEU classifiers regardless of the sample size. Additionally, the McNemar's tests (Table 7) indicated that the differences between the No-RFE and RFE classifications were statistically significant for all NEU classifications, with the exception of classifications trained from 1,250 and 2,500 samples.

While the RF classifier generally provided the highest overall accuracies of all three supervised machine learning algorithms, the RFE process was generally not helpful for that classifier (Figure 6). Indeed, only sample sizes of 40 and 626 saw an increase in RF accuracy after RFE, for the other seven sample sizes, the No-RFE classification had slightly higher performance than the RFE classification (Table 8). Furthermore, the largest difference between any No-RFE and RFE RF classification was the decrease in accuracy of 0.3%, which occurred when RFE was used with the sample size of 159. A possible reason for the decline in accuracy with RFE feature selection is that RF builds an ensemble of decision trees in which each tree uses only a subset of the variables. Apparently, therefore, RF is able to exploit somewhat variables without requiring any additional feature selection. On the other hand, perhaps with a much larger feature set, for example, hundreds or thousands of features, then RFE may be of value for RF. In addition, perhaps other feature selection methods may be more effective with RF.

In summary, RFE provides a statistically significant increase of 4% to 5% for small sample sizes, and for the SVM and NEU classifier. For larger sample sizes, the benefit is smaller, and less consistent. In contrast, for RF, RFE may actually decrease the overall accuracy, even for small training sample sets. Therefore, we recommend as part of best practices supervised machine learning classification of remotely sensed data that feature selection methods should be at least be explored to see if

classifications trained from feature sets with reduced dimensionality provide any improvement to overall map accuracy, especially if training data are limited.

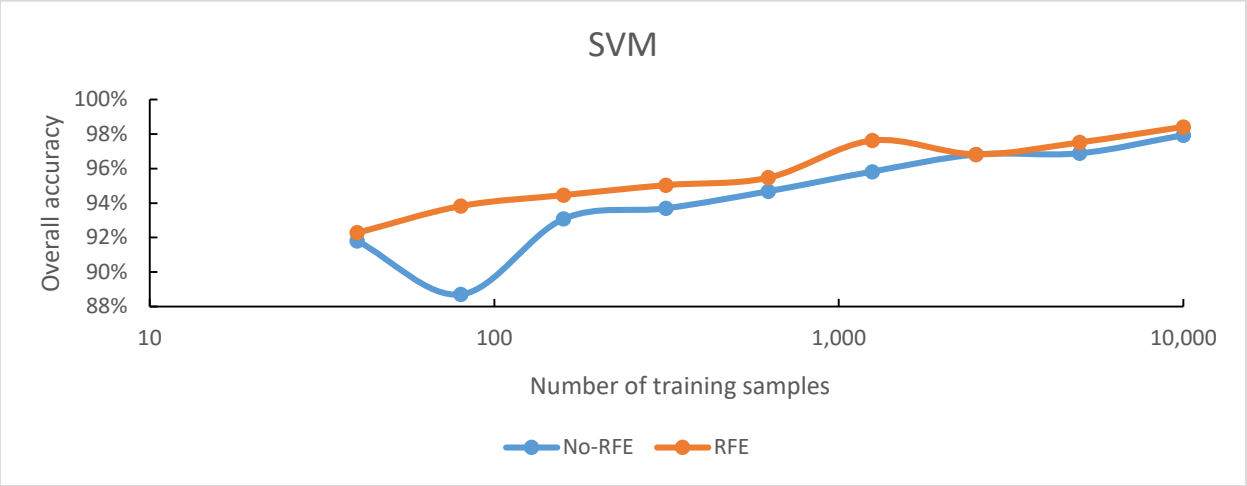

Figure 5 - Overall accuracy of No-RFE and RFE SVM Classifications. (Note that the x-axis is a log scale).

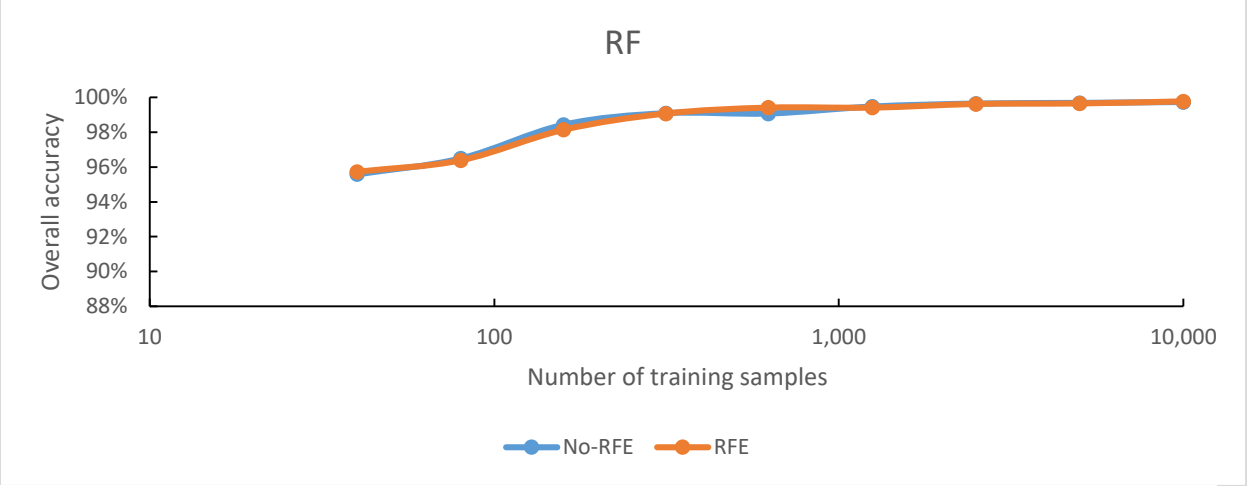

Figure 6 - Overall Accuracy of No-RFE and RFE RF classifications. (Note that the x-axis is a log scale).

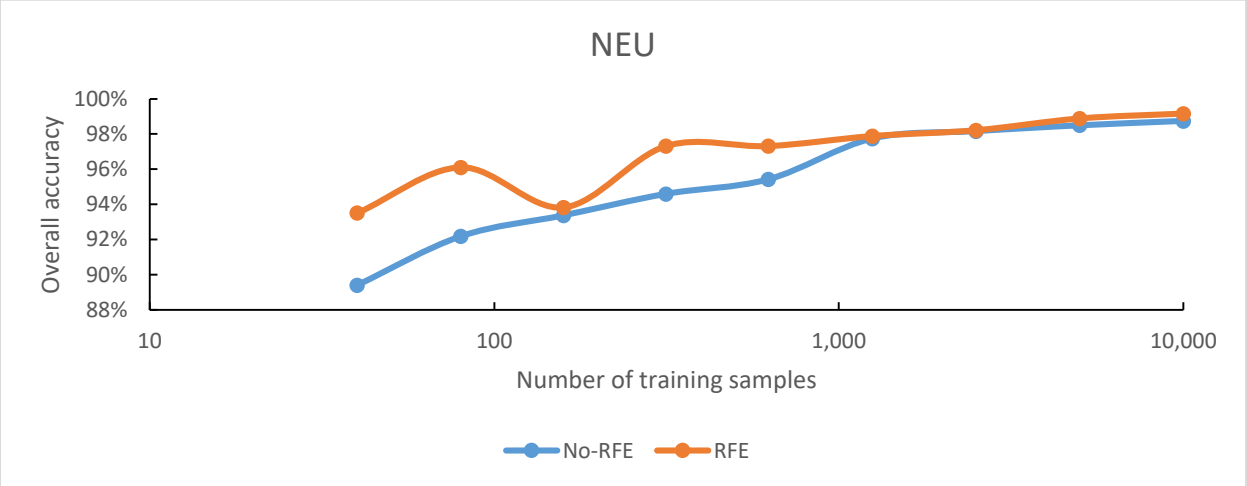

Figure 7 - Overall Accuracy of No-RFE and RFE NEU classifications. (Note that the x-axis is a log scale).

Table 7 – McNemar's test of statistical difference for classification accuracy with and without RFE (indicates differences between classifications that are statistically significant, p < 0.05)*

Number of training samples	SVM		RF		NEU	
	RFE > Non RFE accuracy	Significance of difference (*p*-value)	RFE > Non RFE accuracy	Significance of difference (*p*-value)	RFE > Non RFE accuracy	Significance of difference (*p*-value)
40	Y	< 0.001*	Y	< 0.001*	Y	< 0.001*
80	Y	< 0.001*	N	0.06	Y	< 0.001*
159	Y	0.003*	N	< 0.001*	Y	< 0.001*
315	Y	< 0.001*	N	< 0.001*	Y	< 0.001*
626	Y	< 0.001*	Y	< 0.001*	Y	< 0.001*
1250	Y	< 0.001*	N	< 0.001*	Y	0.3808
2500	N	< 0.001*	N	< 0.001*	N	0.7247
5000	Y	< 0.001*	N	< 0.001*	Y	< 0.001*
10000	Y	< 0.001*	N	< 0.001*	Y	< 0.001*

Table 8 - Confusion Matrix for the NEU classification trained from 40 samples

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified Data (No. Objects)	**Forest**	7263	43	11	6	7323	99.2%
	Grassland	682	1152	32	2	1868	61.7%
	Other	95	60	474	10	639	74.2%
	Water	45	1	73	51	170	30.0%
	Total	8085	1256	590	69	10000	**Overall Accuracy: 89.4%**
	Producer's Accuracy	89.8%	91.7%	80.3%	73.9%		

Table 9 - Confusion Matrix for the NEU-RFE classification trained from 40 samples

		Reference Data (No. Objects)					
		Forest	**Grassland**	**Other**	**Water**	**Total**	**User's Accuracy**
Classified Data (No. Objects)	**Forest**	7750	115	22	17	7904	98.1%
	Grassland	187	1112	91	2	1392	79.9%
	Other	145	29	456	18	648	70.4%
	Water	3	0	21	32	56	57.1%
	Total	8085	1256	590	69	10000	**Overall Accuracy: 93.5%**
	Producer's Accuracy	95.9%	88.5%	77.3%	46.4%		

5. Conclusion

This investigation explored the benefits of recursive feature elimination (RFE) in improving classification accuracy for three commonly used supervised machine algorithms, SVM, RF, and NEU, trained from varying sample sizes in a regional-scale land-cover classification of HR remotely sensed data. The results indicate that feature selection can be beneficial for SVM and NEU. Previous research on the robustness of SVM to high dimensional data has been contradictory. Our findings support those of Pal and Foody (2010), who observed that the SVM classifier may be sensitive to the dimensionality of the dataset as the RFE process usually improved the performance of the SVM classifier. In contrast, for RF, applying RFE feature selection in general resulted in slightly lower accuracies.

The benefit of feature selection for SVM and NEU is strongly dependent on sample size, as would be expected based on the Hughes phenomenon. The largest improvement in accuracy was found for small training sample sizes, and the smallest improvement at the largest sample sizes, 2,500 and greater. It is worth noting, however, that even when the sample size was very large (n = 10,000) the RFE process still improved the overall accuracy of SVM and NEU by 0.5% and 0.4%, respectively. Thus, even with very large samples, it may be beneficial to incorporate feature selection into standard SVM and NEU classification.

When the RFE-optimized feature sets were used in both the SVM and NEU classifiers, the improvement in performance on some classifications were superior to No-RFE SVM and NEU classifications that were trained using much larger sample sets. For example, when the RFE-optimized feature sets were used with the SVM classifier trained from 1,250 samples, the overall accuracy was nearly equivalent to the No-RFE SVM classifier trained from 10,000 samples, even though the SVM classification trained from 10,000 samples contained 8-times the number of samples. Thus, spending

resources on feature selection can result in a greater increase in accuracy than spending those same resources in collecting more reference data.

In summary, feature selection processes such as RFE are a valuable pre-processing step that should be incorporated or explored in remote sensing analyses, especially analyses involving high dimensional datasets, with limited training data. While the RFE process did not prove to be advantageous for the RF classifier in this case, other feature selection methods may be useful for RF. Therefore, even for RF, it may still be a good practice in applied studies to investigate if feature selection processes provide any improvement in overall accuracy even when using a RF classifier. As feature selection is an optional additional step in the classification process, it does not have to be included in the classification if it found to have a negative effect on the overall accuracy of the classifier. For SVM and NEU, feature selection appears to be very important.

References

Alonso, M. C., Malpica, J. A., and Martinez de Agirre, A. (2011). Consequences of the Hughes Phenomenon on Some Classification Techniques. *ASPRS 2011 Annual Conference Milwaukee, Wisconsin, May 1-5*. 1–9.

Archibald, R., Fann, G. (2007). Feature Selection and Classification of Hyperspectral Images With Support Vector Machines. *IEEE Geoscience and Remote Sensing Letters.* 4(4). 674-677. DOI: 10.1109/LGRS.2007.905116.

Arvor, D., L. Durieux, S. Andrés, and Laporte, M.A. (2013). Advances in Geographic Object-Based Image Analysis with Ontologies: A Review of Main Contributions and Limitations from a Remote Sensing Perspective. *ISPRS Journal of Photogrammetry and Remote Sensing,* 82: 125–137. DOI:10.1016/j.isprsjprs.2013.05.003

Baatz, M. and Schäpe, A. (2000). Multiresolution segmentation – an optimization approach for high quality multi-scale image segmentation. Strobl, T. Blaschke, G. Griesebner (Eds.), Angewandte Geographische Informations-Verarbeitung XII, Wichmann Verlag, Karlsruhe, Germany (2000) pp, 12-23.

Belgiu, M. and Drăgut, L. (2014). Comparing supervised and unsupervised multiresolution segmentation approaches for extracting buildings from very high resolution imagery. ISPRS Journal of Photogrammetry and Remote Sensing. 96: 67-75. https://doi.org/10.1016/j.isprsjprs.2014.07.002.

Bittencourt, H. R., and Clarke, R. T. (2003). Use of Classification and Regression Trees (CART) to Classify Remotely-Sensed Digital Images. *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No.03CH37477)* July 21-25, 2003. Toulouse, France. DOI: 10.1109/IGARSS.2003.1295258.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Belmont, CA: Wadswoth, 1984.

Colkesen, I., and Kavzoglu, T. (2016). Performance evaluation of rotation forest for svm-based recursive feature elimination using hyperspectral imagery. *8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2016.* Los Angeles, CA. Aug. 21-24, 2016. DOI: 10.1109/WHISPERS.2016.8071792.

Drăgut, L., Csillik, O., Eisank, C., and Tiede, D. (2014). Automated parameterization for multi-scale image segmentation on multiple layers. ISPRS J. Photogrammetric Remote Sensing. 88(100): 119-127. https://doi.org/10.1016/j.isprsjprs.2013.11.018.

Foody, G. M. (2017). Impacts of Sample Design for Validation Data on the Accuracy of Feedforward Neural Network Classification. *Applied Sciences*. 7(9), 888. https://doi.org/10.3390/app7090888.

Guan, H., Ji, Z., Zhong, L., Li, J., and Ren, Q. (2013). Partially supervised hierarchical classification for urban features from lidar data with aerial imagery. *International Journal of Remote Sensing*. 34(1). 190-210. https://doi.org/10.1080/01431161.2012.712228.

Guo, Q., Kelly, M., Gong, P., and Liu, D. (2007). An Object-Based Classification Approach in Mapping Tree Mortality Using High Spatial Resolution Imagery." *GIScience & Remote Sensing*. 44(1): 24-47. DOI: 10.2747/1548-1603.44.1.24.

Hay, G. J., Castilla, G., Wulder, M.A., Ruiz, J.R. (2005). An automated object-based approach for the multiscale image segmentation of forest scenes. International Journal of Applied Earth Observation and Geoinfromation. 7(4): 339-359. DOI: 10.1016/j.jag.2005.06.005.

Huang, M., Hung, Y., Lee, W. M., Li, R. K., Jiang, B. (2014). SVM-RFE Based Feature Selection and Taguchi Parameters Optimization for Multiclass SVM Classifier. *Scientific World Journal*. 2014 (795624). http://dx.doi.org/10.1155/2014/795624.

Huang, Y., Zhao, C., Yang, H., Song, X., Chen, J., and Li, Z. (2017). Feature Selection Solution with High Dimensionality and Low-Sample Size for Land Cover Classification in Object-Based Image Analysis. *Remote Sensing.* 9(2017), 939-955. doi:10.3390/rs9090939.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory.* 14(1). 55-63. DOI: 10.1109/TIT.1968.1054102.

Imani, M. and Ghassemian, H. (2015). Feature reduction of hyperspectral images: Discriminant analysis and the first principle component. *Journal of AI and Data Mining.* 3(1). 1-9. doi:10.5829/idosi.JAIDM.2015.03.01.01.

Jović, A., Brkić, K., and Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 25-29 May 2015, Opatija, Croatia. DOI: 10.1109/MIPRO.2015.7160458.

Kaushik, S. (2016). Introduction to Feature Selection methods with an example (or how to select the right variables?) *Analytics Vidhya.* https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/.

Kavzoglu, T., and Tonbul, H. (2018). An experimental comparison of multi-resolution segmentation, SLIC, and K-means clustering for object-based classification of VHR imagery. *International Journal of Remote Sensing*. 39(18). 6020-6036. https://doi.org/10.1080/01431161.2018.1506592.

Kim, M., Warner, T.A., Madden, M., and Atkinson, D. (2011). Multi-scale texture segmentation and classification of salt marsh using digital aerial imagery with very high spatial resolution. International Journal of Remote Sensing. 32: 2825-2850.

Kuhn, Max. (2016). caret: Classification and Regression Training. R package version 6.0-71. https://CRAN.R-project.org/package=caret.

Lear, R.F. (2005). NAIP Quality Samples. United States Department of Agriculture Aerial Photography Field Office. Available Online: https://www.fsa.usda.gov/Internet/FSA_File/naip_quality_samples_pdf.pdf (last date accessed: 28 Dec 2018).

Maxwell, A.E., Strager, M. P., Warner, T. A., Zegre, N.P., Yuill C. B. (2014). Comparison of NAIP orthophotragephy and RapidEye satellite imagery for mapping of mining and mine reclamation. GIScience & Remote Sensing. 51(3):301-320. https://doi.org/10.1080/15481603.2014.912874.

Maxwell, A. E., T.A. Warner, B.C. Vanderbilt, and Ramezan, C.A. (2017). Land cover classification and feature extraction from National Agriculture Imagery Program (NAIP) Orthoimagery: A review. Photogrammetric Engineering and Remote Sensing 83(11): 737-747. DOI: 10.14358/PERS.83.10.737.

Melgani, F. and Bruzzone, L. (2004) Classification of Hyperspectral Remote Sensing Images with Support Vector Machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42, 1778-1790. https://doi.org/10.1109/TGRS.2004.831865

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 12(2): 153-157. DOI:10.1007/BF02295996.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C., Lin, C. (2012). Support Vector Machines: The Interface to libsvm in package e1071. R package version 6.0-71. https://CRAN.R-project.org/package=e1071. (last date accessed: 18 Feb 2019).

Pal., M., and Foody, F. M. (2010). Feature Selection for Classification of Hyperspectral Data. *IEEE Transactions on Geoscience and Remote Sensing.* 48(5). 2297-2307. doi:10.1109/TGRS.2009.2039484.

Petrie, G and Toth, C. K. (2008). Airborne and Spaceborne Laser Profilers and Scanners In Shan, Jie and Toth, Charles K. *Topographic Laser Ranging and Scanning: Principles and Processing*. CRC Press.

Ramezan, C. A., Warner, T. A., Maxwell A. E. (2019a). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sensing. 11(2): 185. https://doi.org/10.3390/rs11020185.

Ramezan, C. A., Warner, T. A., Maxwell A. E. (2019b). What is the Optimal Training Sample Size for Common Machine Learning Classifiers? In review by *Photogrammetric Engineering and Remote Sensing*.

Ripley, B., and Venables, W. (2016). Package 'nnet': Feed-Forward Neural Networks and Multinomial Log-Linear Models. R package version 7.3-12.

Stevens, A., Nocita, M., Tóth, G., Montanarella, L., and van Wesemael., B. (2013) Prediction of Soil Organic Carbon at the European Scale by Visible and Near InfraRed Reflectance Spectroscopy. *PLoS ONE* 8(6): e66409. https://doi.org/10.1371/journal.pone.0066409.

Thenkabail, P. S., Gumma, M. K., Teluguntla, P., and Mohammed, I. A. (2014). Hyperspectral Remote Sensing of Vegetation and Agricultural Crops. *Photogrammetric Engineering and Remote Sensing*. 80(8). 692-709.

Van Campenhout, J. M. (1978). On the peaking of the Hughes mean recognition accuracy: the resolution of an apparent paradox. *IEEE Transactions on Systems, Man, and Cybernetics*. 8(5): 390-395. DOI: 10.1109/TSMC.1978.4309980

Warner, T., Steinmaus, K. and Foote, H. (1999). An evaluation of spatial autocorrelation-based feature selection. *International Journal of Remote Sensing* 20 (8): 1601-1616. DOI: 10.1080/014311699212632

Wright, M. N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*. 77. 1-17. DOI: 10.18637/jss.v077.i01.

WVU NRAC. (2013). Aerial Lidar Acquistion Report: Preston County and North Branch (Potomac) LIDAR *.LAS 1.2 Data Comprehensive and Bare Earth. West Virginia Department of Environmental Protection. Available online: http://wvgis.wvu.edu/lidar/data/WVDEP_2011_Deliverable4/WVDEP_deliverable_4_Project_Report.pdf (accessed on 1 December 2018).

Yu, S. (2003). Feature Selection and Classifier Ensembles: A Study on Hyperspectral Remote Sensing Data. 1-128. https://pdfs.semanticscholar.org/7728/bf4d5689f49fc45571b4c90acebe55da541c.pdf

Zhang, R., and Ma, J. (2009). Feature selection for hyperspectral data based on recursive support vector machines. *International Journal of Remote Sensing* 30 (14). 3669-3677. https://doi.org/10.1080/01431160802609718.

Chapter 5 – Conclusion

1. Overall findings of this research

This dissertation focused on the development of regional-scale object-based land-cover classifications using high spatial resolution (HR) (1 – 5 m) remotely sensed data (aerial multispectral imagery, and light detection and ranging LIDAR point cloud data). The aim of this dissertation was to examine several core methodological processes in remote sensing analyses, such as training sample selection, cross-validation parameter tuning (Chapter 2), supervised machine learning algorithms, training set size (Chapter 3), and feature selection (Chapter 4) on a large, regional-scale HR dataset. These processes were examined through a series of three related experiments. The regional-scale study area for all three papers was approximately 260,975 ha, which is notably larger than most methodological and applied object-based, or geographic object-based image analyses (GEOBIA) using HR remote sensed data (Ma et al., 2017).

Chapter 2's findings indicate that for support vector machines (SVM) classification, training sets acquired through statistical stratified random sampling methods provide higher levels of overall accuracy than training sets acquired via simple random, or deliberative sampling methods. There were also minimal differences in overall accuracy for SVM classifications tuned using three different cross-validation parameter tuning methods: *k*-fold, Monte Carlo, and leave one out. However, Monte Carlo and leave one out greatly increased the processing time of the SVM classification. As a result, *k*-fold cross-validation was found to be preferable, especially if processing and time resources are limited. Additionally, SVM classifications trained from sample sets collected from a relatively small subset area of the regional-scale dataset had similar levels of overall accuracy to classifications trained from sample sets of equivalent size collected across the regional-scale area. As training data collection can be expensive over large areas, subset-

based sample selection can be advantageous for reducing sampling costs, especially if field observations are needed.

Chapter 3 investigated the effect of training set size on five supervised machine learning classifiers: SVM, random forest (RF), *k*-nearest neighbors (*k*-NN), single layer perceptron neural networks (NEU), and learning vector quantization (LVQ) on regional-scale HR analyses. The results demonstrated that each supervised classifier responded differently to variations in sample size, although larger training sets generally improved overall accuracy, a similar finding to Foody et al. (2006). RF consistently outperformed the other four classification algorithms when trained from any sample size. RF and LVQ also quickly plateaued in performance when training size reached (*n* = 315), however LVQ was also the worst performing classifier out of all five methods. The overall accuracy of SVM and NEU classifiers continued to increase with larger training sets, even when the sample size reached a very large size (*n* = 10,000). However, the processing time of the NEU classifier was extremely high with very large sample sets (i.e. *n* = 5,000, 10,000). *K*-NN was found to have the shortest processing time of all algorithms, but may be more sensitive to the characteristics, rather than the size of the training set. Thus, determining an optimal training set size can vary on the classification method used for analysis. Furthermore, if training data are limited, classifiers such as RF may be preferable, as RF was able to produce superior or equivalent levels of accuracy to other classification methods, even when trained from much larger sample sets.

Chapter 4 investigated how feature selection using recursive feature elimination (RFE) applied to datasets with variable sized training sets affected the overall accuracy of three supervised machine learning algorithms: SVM, RF, and NEU. The results of this analysis suggested that RFE was generally advantageous for improving the overall accuracy of both SVM and NEU classifiers, however the benefits to overall accuracy diminished as sample size

increased. RFE was not very effective in improving the overall accuracy of the RF classifier when trained from both very small and very large sample sets, which suggests that RF is more robust to the Hughes phenomenon that SVM or NEU classifiers. Feature selection methods such as RFE are optional, and do not have to be incorporated into the final classification if feature selection did not improve the overall accuracy of the classifier. Nevertheless, feature selection does in many cases increase the overall classification accuracy. Therefore, we recommend that feature selection be considered part of best practices when conducting remote sensing classifications, especially on high dimensional datasets with limited training data.

2. Limitations and technical comments

While the size of the HR regional-scale dataset examined in this dissertation (260,975 ha) was far larger than HR datasets used in most GEOBIA analyses (Ma et al., 2018), the size of the study area was limited by the available LIDAR data. While a larger scale analysis would have been possible by using only NAIP data, it was decided that the use of LIDAR to provide additional spectral, and elevation information to assist with the classification outweighed the limitations on the extent of the study area. As seen in Chapter 4, LIDAR derived variables, especially LIDAR intensity variables, were highly ranked by the feature selection process. Similar to the findings of Kashani et al. (2015), Song et al. (2002), and Maxwell et al. (2015), LIDAR intensity was found to be valuable for land-cover classification. However, it should be mentioned that for future regional-scale land-cover analyses, large area LIDAR data may not always be available. While NAIP data are typically available on a state-wide level (Maxwell et al., 2017), large-area or state-wide LIDAR datasets are rare when compared to the availability of large, regional-area HR multispectral imagery.

Regarding the processing of large, regional-scale HR land-cover datasets, the immense size of both the NAIP and LIDAR datasets caused several technical challenges, especially during the image processing and image segmentation stages. While software platforms such as ArcGIS (ESRI, 2017), and ERDAS Imagine (Hexagon, 2018) are built to handle large geospatial datasets, hardware resource limitations, such as system memory availability led to processing bottlenecks. For example, the rasterization of the LIDAR point cloud into intensity and elevation-based rasters took several days of computing. The color balancing (Lear, 2005) of the NAIP orthomosaic also required several days to process. Although ArcGIS Pro (ESRI, 2019) does incorporate parallel processing in the form of multithreading, and this can decrease time costs with processing large datasets, processing regional-scale HR datasets can still be expensive in terms of time, and requires a large computing resources.

An additional concern that affects analyses conducted in a GEOBIA framework was the image segmentation process, which was costly in terms of processing power and time when conducted on regional-scale HR datasets. While the multi-resolution segmentation (MRS) (Baatz and Schape, 2002) algorithm within the eCognition Developer (Trimble, 2018a) was able to successfully segment the regional-scale dataset, the segmentation took several days to complete, and required at least 16 GB of memory. Due to the hardware limitations of a single system, we recommend future analyses involving large, regional-scale HR datasets explore server, or cluster-based processing options, which may remediate some of the issues encountered through processing on a single workstation. Several of the software suites used in this analysis, have server or cloud-based processing capabilities, such as ArcGIS Server (ESRI, 2018), and eCognition Server (Trimble, 2018b), which can be advantageous for processing very large remotely sensed datasets (Yan et al., 2017). Future hardware advances in information technology may also mitigate some of the processing difficulties encountered in this analysis.

Additionally, alternate image segmentation algorithms such as the multi-threshold segmentation (Li and Shao, 2014) may be less expensive in terms of processing time and resource demands than the MRS algorithm.

3. Conclusions and Recommendations

The results of the analyses in this dissertation provided several insights on developing land-cover classifications via supervised machine learning classifiers on large, regional-scale HR remotely sensed datasets.

When designing a regional-scale HR land-cover classification, the choice of supervised machine learning algorithm for classification is important, as accuracy can vary between different supervised classifiers when applied to classify the same dataset (Maxwell et al., 2018; Noi and Kappas, 2018; Raczko and Zagajewski, 2017). Similar to the observations made by Qian et al., (2015), in this study, the supervised classifiers responded in different ways to variations in training set size. Different supervised machine learning algorithms also varied in sensitivity to the Hughes phenomenon (Hughes, 1968). The negative effects of the Hughes phenomenon on classification accuracy can be a particular concern if analyzing high dimensional datasets such as hyperspectral data, which are becoming increasingly available. As training data may not be abundant or very expensive to collect due to the size of regional-scale study areas, it may be advantageous to select a classifier such as random forests (RF), which in this case was robust to the Hughes phenomenon, and was able to provide high levels of overall accuracy even when trained from limited training sets. Several other studies have also highlighted the strengths of the RF classifier (Ham et al., 2005; Maxwell et al., 2018). Additionally, on large, homogenous study areas, sampling from relatively small subset areas of the regional-scale dataset can be effective in reducing training data collection costs, provided the sampling area contains adequate examples of the land-cover classes of interest.

Overall, as this dissertation has demonstrated, supervised machine learning algorithms can be used to develop object-based land-cover classifications of large, regional-scale HR remotely sensed data. However, when designing a land-cover classification or analysis, all aspects of the classification process, such as sampling design or classifier selection should be carefully considered with respect to the dataset of interest and study objectives. As this work is, as far as I am aware, the first of its kind to examine several core remote sensing classification processes on large, regional-scale HR datasets, hopefully future HR object-based regional-scale analyses and land-cover classifications can use this dissertation as a guide towards providing high quality regional-scale classifications of HR remotely sensed data.

References

Baatz, M.; Schäpe, A. (2002). Multiresolution segmentation—An optimization approach for high quality multi-scale image segmentation. In Proceedings of the Angewandte Geographische Informations-Verarbeitung XII, Wichmann Verlag, Karlsruhe, Germany, 2000; pp. 12–23.

ESRI. (2017). ArcGIS Desktop: Release 10.5.1; Environmental Systems Research Institute: Redlands, CA, USA, 2017.

ESRI. (2018). ArcGIS Server; Environmental Systems Research Institute: Redlands, CA, USA, 2018.

ESRI. (2019). ArcGIS Pro: Release 2.3.0; Environmental Systems Research Institute: Redlands, CA, USA, 2019.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. Remote Sensing of Environment. 1(15): 1-14. https://doi.org/10.1016/j.rse.2006.03.004.

Hexagon AB. (2018). ERDAS IMAGINE 2018; Hexagon Geospatial: Madison, AL 35758, USA, 2018.

Hughes, G. (1968). On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory. 14(1). 55-63. DOI: 10.1109/TIT.1968.1054102.

Kashani, A.G.; Olsen, M.; Parrish, C.; Wilson, N. (2015). A Review of LIDAR Radiometric Processing: From Ad Hoc Intensity correction to Rigorous Radiometric Calibration. *Sensors.* 15, 28099–28128, doi:10.3390/s151128099.

Lear, R.F. (2005). NAIP Quality Samples. United States Department of Agriculture Aerial Photography Field Office. Available Online: https://www.fsa.usda.gov/Internet/FSA_File/naip_quality_samples_pdf.pdf (last date accessed: 28 Dec 2018).

Li, Xiaoxiao and Shao, Guofan (2014). Object-Based Land-Cover Mapping with High Resolution Aerial Photography at a County Scale in Midwestern USA. Remote Sensing. 6: 11372-11390. doi:10.3390/rs61111372.

Ma, L.; Li, M.; Ma, X.; Cheng, K.; Du, P.; Liu, Y. (2017). A review of supervised object-based land-cover image classification. ISPRS J. Photogramm. Remote Sensing. 130, 277–293, doi:10.1016/j.isprsjprs.2017.06.001.

Maxwell, A.E.; Warner, T.A.; Strager, M.P.; Conley, J.F.; Sharp, A.L. (2015). Assessing machine learning algorithms and image- and LiDAR-derived variables for GEOBIA classification of mining and mine reclamation. Int. J. Remote Sensing. 36, 954–978, doi:10.1080/01431161.2014.1001086.

Noi, P. T., Kappas, M. (2018). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors. 18(1): 18. https://doi.org/10.3390/s18010018.

Qian, Y., Zhou, W., Yan, J., Li, W., Han, L. (2015). Comparing Machine Learning Classifiers for Object-Based Land Cover Classification Using Very High Resolution Imagery. Remote Sensing. 7(1): 153-168. https://doi.org/10.3390/rs70100153.

Raczko, E., and Zagajewski, B. (2017). Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. European Journal of Remote Sensing. 50(1): 144-154. https://doi.org/10.1080/22797254.2017.1299557.

Song, J.H.; Han, S.H.; Yu, K.Y.; Kim, Y.I. (2002). Assessing the possibility of land-cover classification using LIDAR intensity data. Int. Arch. Photogramm. Remote Sensing Spatial Information Science. 34, 259–262.

Trimble. (2018a) Trimble eCognition Suite 9.3.2; Trimble Germany GmbH: Munich, Germany, 2018.

Trimble. (2018b) Trimble eCognition Server 9; Trimble Germany GmbH: Munich, Germany, 2018.

Yan, J., Ma, Y., Wang, Kim-Kwang, R. C., Jie, W. (2017). A cloud-based remote sensing data production system. Future Generation Computer Systems. 86(2018). 1154-1166. DOI: 10.1016/j.future.2017.02.044.